# THE DESIGN PRINCIPLES OF AN ONLINE LISTENING TEST FOR INVESTIGATING THE PERCEPTION OF HEAVY METAL HARMONY

Otso Björklund

Master's thesis

University of Helsinki
Department of Philosophy, History,
Culture, and Art Studies
Musicology

April 2015

TABLE OF CONTENTS

# 1. INTRODUCTION

This study describes the design principles of a large scale online listening test for investigating the perception of heavy metal harmony. The large scale online listening test designed in this study is to be used in Dr. Esa Lilja's research project on the subject. As online listening tests are still a fairly new method for investigating music perception, this whole study is dedicated to investigating the methodology. The main research question is how to investigate the perception of heavy metal harmony using an online listening test. The terms *Internet experiment* and *online experiment* are used interchangeably in this study to refer to an experiment that the subjects participate in using a computer with an Internet connection in any location at any suitable time.

Listening tests are used in various fields of research ranging from audio quality testing to cognitive science. This study focuses on listening test methodology as it relates to the study of music perception. Experiments on the perception of audio quality and the technical details of creating an online listening test application are not described in this study. Although many of the fundamentals of listening test methodology are practically the same regardless of the field of research, some methodology is borrowed from the field of perceptual audio quality evaluation.

The different aspects of music perception have been investigated with various listening test experiments. The main purpose of such experiments has typically been to develop models of music perception and to test whether the traditional concepts of music theory correspond to the perception of music. One of the central figures of research into the perception of harmony is Carol Krumhansl, who has investigated the subject for decades. Some of the most seminal studies in which she has been involved in are described in the third chapter of this study, along with some recent studies on the same subject (e.g. McDermott et al. 2010). Essentially, the purpose of these studies has been to investigate what people find consonant or dissonant and how different pitches and chords are perceived to fit into different contexts.

Lilja's (2009) doctoral dissertation *Theory and Analysis of Classic Heavy Metal Harmony* is possibly one of the foremost studies on the subject of heavy metal harmony. Potentially the most important argument Lilja (ibid: 211) makes in his dissertation is that the effects of distortion change the harmonic content of chords to such an extent that they no longer correspond to their notated forms, which makes it necessary to reconsider the concepts of consonance and dissonance in heavy metal music. The purpose of the large scale experiment, whose design principles this study describes, is to investigate which intervals are considered most consonant and which least consonant in relation to the *power chord* (cf. ibid: 102). According to Lilja (ibid) "[t]he power chord is an especially normative musical feature for heavy metal", which implicates that the investigation of the perception of consonance in relation to the power chord is an essential part of investigating the perception of heavy metal harmony. In essence, the purpose of this study is to design an experiment that, to some extent, follows in the footsteps of earlier music perception experiments but instead of the laboratory the design presented here is for an experiment conducted on the Internet. The results of the experiment designed in this study can be compared to the results of earlier studies on the perception of non-distorted harmony to see how distortion affects the perception of harmony.

The basic methodology of listening tests is reviewed (Chapter 2) along with numerous studies on music perception (Chapter 3) to find the best practices of music perception experimenting. Studies involving Krumhansl on the perception of harmony are analysed in detail as they offer insight into the practices of experiments on the perception of harmony. The methodology of Internet experimenting is also reviewed, and some online listening test experiments on music perception are described (Chapter 4). Online listening tests for investigating different aspects of the perception of rhythm have been used by Wright (2008) and Honing (cf. chapter 4.3. of this study).

A pilot experiment was conducted to test three different experimental methods (Chapter 5). These methods were selected from the literature on the methodology of listening tests and from the experiments involving Krumhansl (e.g. Krumhansl & Kessler 1982). The purpose of the pilot experiment was to gather data on different experimental

methods in order to select a suitable method for a large scale online listening test. The results of the pilot experiment were also used to improve the selected method.

The second chapter of this study describes the relevant basic methodology of listening tests and the potential issues of experimenting, such as bias. Some general aspects of sensory testing are described in relation to listening tests. In the third chapter many classic experiments on pitch and music perception are described and analysed. The purpose of the third chapter is to offer an overview of earlier studies and their experimental practices. The fourth chapter introduces the methodology of Internet experimenting and some recent online listening tests on music perception are described. The potential issues of Internet experimenting are also discussed. The methodology and results of the pilot experiment are presented in chapter five. The sixth chapter discusses the application of the pilot experiment's results to my design for a large scale online experiment. A user interface was designed based on suggestions made by the subjects who participated in the pilot experiment. The appendices A1 and A2 contain the materials of the pilot experiment and its results.

# 2. BASIC METHODOLOGY OF LISTENING TESTS

Listening tests are used in various areas of research: music psychology, cognitive science, audio quality research and psychophysics just to name a few. Listening tests share much of their methodology with the sensory evaluation methods used in the sensory evaluation of food. Regardless of the field, there are many aspects and considerations which are common to all listening tests and other tests using sensory evaluation as a method of measurement. This chapter describes the fundamentals of listening tests. I will refer to a listener participating in a listening test as *subject* or *assessor,* and I will use *sample* or *stimulus* to denote the samples that are presented to the subject during a test.

## 2.1. Defining the response attribute and task

Measuring the human perception of auditory events directly is not possible (Bech & Zacharov 2006: 1). By asking the listener to quantify their experience it is possible to measure some aspects of the perception of an auditory event. According to Bech and Zacharov (ibid), this kind of evaluation is often the basis of a listening test. The experimenter can only access the experience of the listener via the listener's description of the experience, and the goal of a listening test experiment is to find a mapping from the physical sound event to the auditory experience of the listener (Hynninen 2001: 3–4).

The most essential considerations in designing a listening test, just like in all experiments, are defining the research question and hypothesis (Bech & Zacharov 2006: 17). Having defined these, it is possible to begin the design process. In order to quantify the experiences of the subject, it is necessary to define the dependent variable, which is represented by the subject’s answers. There are also independent variables, which the experimenter controls (ibid: 98). The attribute that is being investigated is called the *response attribute* and the method of quantifying this attribute by the assessor is called *response format*. In the case of complex stimuli, the process of defining the response

attributes is the following: first the relevant attributes of the stimuli should be identified, then the attributes should be described and finally these attributes will be evaluated in the actual experiments. (Ibid: 39–40.)

As an example of these terms, let us consider an imaginary example. In an experiment used to investigate the effects of tempo on the irritativeness of a musical sample, the response attribute could be irritativeness and the dependent variable could be the irritativeness measured on some scale defined by the response format. The response format could be a rating from one to ten, ten being most irritating. Such an experiment would require the samples to be identical in every other aspect except for tempo.

Once the research question and dependent variable have been defined, the practical aspects of the test can be considered. Depending on the selected method of testing, the test can be very time-consuming for the subjects. Bech and Zacharov (2006: 302) consider 20 to 40 minutes to be a suitable length for a session of testing. If the entire test requires more time from a subject, it will be best to divide the test into multiple sessions to avoid fatiguing the subject. One possible way of doing this is using an incomplete block design which consists of several sessions called blocks, each block containing a subset of all the samples (Lawless & Heymann 1998: 105). This way the subject can evaluate all of the samples during the experiment but in multiple sessions. If the number of samples to be tested is not too large, a complete block design can be utilised. In a complete block design, all of the samples are presented to the subject in a single session (Lawless & Heymann 1998: 806).

The subjects should be presented with a task that is easy to perform and easy to understand. The process of informing the subjects of the task and instructing them how to perform it without difficulties is called familiarisation (Bech & Zacharov 2006: 310–311). Bech and Zacharov (2006: 311–315) divide familiarisation into multiple steps that consist of introducing the task, instructing the subject either verbally or textually how to perform the task, and familiarising the subject with the task's user-interface and samples.

## 2.2. Measuring experiences and impressions

Measuring the impressions of subjects requires the use of some scaling method. In sensory testing, the objective of the test affects the selection of the scale used for measuring the response attribute. The selection of subjects and the kind of samples being tested also need to be considered when selecting a scaling method for a sensory test (Stone et al. 2012: 83). The same factors must be considered in the selection of a scaling method for a listening test. The purpose of the measurements must also be considered. Measurements can be very specific, aiming to measure the magnitude of a single attribute, or they can aim to measure an overall impression of the stimulus (Bech & Zacharov 2006: 3).

Ekman and Sjöberg (1965: 452) raise an important question: "What is being measured?" A question which they think is not raised as often in regard to scales as expected. They also note that there are opposing views as to what a scale measures. According to these opposing views, a scale can reflect a subjective variable well, or it can be arbitrary and even obscure important features of data (cf. ibid). Assuming that scales can be used to measure impressions, the characteristics of well-functioning scaling methods must be considered. Two very important characteristics of scales are identified by Stone et al. (2012: 83–84): scales must be meaningful to the subjects and easy to use.

Scales can be divided into five categories: nominal, ordinal, interval, ratio and multidimensional scales (Hynninen 2001: 4–5). Nominal scales are used for labelling classes. With nominal scales the classification of items and responses can be done with numbers, letters or other symbols (Stone et al. 2012: 86). Hynninen (2001: 4) uses the classification of a sound sample as either male or female voice as an example of nominal scale use. Ordinal scaling is used to put the samples in some order (ibid). The magnitude of difference between samples is not necessarily measured if ordinal scales are used. A very typical type of ordinal scale is ranking. (Stone et al. 2012: 89.) Interval scales do not measure the absolute magnitude of the attribute, although they do measure the difference of magnitude between samples. The scale consists of points which are separated by equal distances (ibid: 96). Ratio scales use the ratio between samples as a

unit of measurement. Hynninen (2001: 4) also includes an example of the use of a ratio scale: the difference in loudness is measured by doubling or halving the loudness. Multidimensional scaling employs a combination of simple scales and it requires the subject to rate the stimuli on multiple scales (Hynninen 2001: 5). These measurements are used to place the samples in a multidimensional space.

Placing samples on an ordinal scale can be done in a variety of ways. In many of these methods, the samples are compared to each other. According to Lawless and Heymann (1998: 301), humans are better at making relative than absolute judgements. One of the simplest alternatives is the use of preference ranking, which is also known as rank order paradigm. The subjects are simply asked to rank the samples in order of preference. One of the benefits of this method is its simplicity. Preference ranking does not require the subject to remember previous samples, and the samples are in a clear frame of reference, as they are presented simultaneously. This is also a disadvantage as samples can only be compared with the samples of the same set. (Lawless & Heymann 1998: 444.) Preference ranking can also be fatiguing if a large number of samples is being ranked because all of the samples must be evaluated before making any judgements (Hynninen 2001: 13).

Another possible method for ranking samples is the method of paired comparison. According to David (1988: 1), this method was popularised by Thurstone (cf. Thurstone 1927a and 1927b). The subject is presented with two samples and instructed to choose the preferred one. In this case preference can also mean that the subject considers the sample to contain more or less of some property. (David 1988: 1.) Once the subject has judged all the possible pairs of samples, an order of preference can be established. David (1988: 2) considers paired comparisons to have two advantages over preference ranking. Paired comparison allows for a more detailed judgement of samples, and it is also easier for the subject if the differences between the samples are not obvious. One of the potential problems with paired comparisons is the possibility of *circular triads* (Kendall & Babington Smith 1940: 327). A circular triad occurs when a subject judges samples inconsistently, for example preferring sample A to B, B to C, and C to A (ibid: 327). These kind of inconsistencies can be a result of samples that are very similar in the

evaluated property, a result of asking the subjects to evaluate a property that is not a linear variable, or the inconsistencies can result from the subject's lack of skill in making reliable judgements (ibid: 325). Kendall and Babington Smith (1940) also present various statistical methods for handling the inconsistencies in paired comparison data. A third possibility for evaluating samples in a reference to each other is triadic comparison. The subject selects the pair that is most alike and the pair that is least alike from a set of three samples (Levelt et al. 1966: 164).

Samples can also be rated individually using suitable scales. A statement about the sample can be presented to the subject and the subject makes the judgement by using a Likert scale, which typically consists of five alternatives ranging from strong agreement to strong disagreement (Lawless and Heymann 1998: 509). The 9-point hedonic scale is very similar to the Likert scale as it also has a neutral centre point and its extremities are the same as on a Likert scale (cf. Stone et al. 2012: 101). Other types of scales can also be used to rate samples. For example Carol Krumhansl (cf. Krumhansl and Shepard 1979) has used seven point rating scales very often in her research, which will be described in more detail in the third chapter of this study. Rating scales typically employ some kind of anchor points such as verbal labels (French-Lazovik and Curtis 1984: 49). Anchor points can be used to define the ends of a scale. For example, the varying degrees of agreement in the Likert scale are anchor points. The purpose of these anchor points is to help the subject use the scale. Anchor points must be selected carefully as they can affect the ratings (ibid: 53).

In the field of sensory evaluation, Stone et al. (2012: 85) consider statistical analysis to be necessary in finding the relationship between the variable controlled by the experimenter and the resulting experiences of subjects. This also applies to listening tests. Different scaling methods offer different possibilities for statistical analysis, which must be taken into account when selecting a scale for a test. These methods of statistical analysis will not be covered in this study as they are too numerous. The subject of statistical analysis is covered in various textbooks (cf. e.g. Nummenmaa 2011).

## 2.3. Test environment in laboratory-based tests

Listening tests are often conducted in specific listening rooms or booths. The listening room has a significant effect on tests which employ loudspeakers, but the effect is not as significant when using headphones. In order to enable repeating a test in another facility, it is necessary to regulate the acoustic properties of the listening space. (Bech & Zacharov 2006: 228.) The effects of the test environment in traditional laboratory based listening tests will only be described briefly as this study is more concerned with online testing.

Background noise can affect the results of listening tests. According to Novitski (2006: 15), noise may have adverse effects on auditory perception. Novitski (2006: 57) shows that the noise of an fMRI (*functional magnetic resonance imaging*) machine has a strong effect on the event-related potential measured during auditory tasks. Practically this means that when measurements of brain activity are taken using an fMRI machine in listening tests, the effects of noise will be visible in the measured reactions to auditory stimuli. Because of the effects that listening room properties and noise have on testing, many technical standards for listening tests contain specifications for listening rooms and background noise levels. These standards often describe the acoustic properties of suitable listening rooms in detail (cf. e.g. ITU-R 2014: 13–14).

Visual cues in the listening test environment can also affect the subject's judgements. Toole and Olive (1994) investigated how knowledge of the model of the loudspeaker affected the subject's judgements of audio quality. Their conclusion is that knowing the model of the loudspeaker has a strong effect on both experienced and inexperienced subjects (Toole & Olive 1994: 15–16). The experimenter must make sure that the subjects are not distracted by any visual cues, and that all of the subjects participating in the experiment have similar visual experiences to ensure the reliability of judgements. The effects of the test environment in Internet experiments are described in the fourth chapter of this study.

## 2.4. Test subjects

In listening tests the subjects have the vital role of making the judgments on the stimuli. Bech and Zacharov (2006: 106–112) list various considerations which affect the selection of subjects for audio quality testing. Many of these considerations apply to most listening tests. Subjects should typically have normal hearing, be available for testing and have the ability to make judgements on the samples (Bech & Zacharov 2006: 107).

The selected subjects should form a representative sample of some population. It is necessary to define the population that the results of the test should apply to. For example, if the objective of a listening test is to investigate how professional musicians experience some aspect of music, the subjects should be selected from the population of professional musicians. Subjects can be either pre- or post-selected. The purpose of pre-selection is to select subjects that belong to some desired part of the whole population and have the abilities the experiment requires (Bech & Zacharov 2006: 118–122). There are various ways to test the subjects' skills for experiments on audio quality and music perception. Pre-selection for audio quality testing can be done using discrimination tests (cf. Bech & Zacharov 2006: 126), and pre-selection for music perception tests can be done based on musical skills, which can be tested using recognition tests (cf. Wallentin et al. 2010). The goal of post-selection is to remove data from subjects whose answers are considered useless (Bech & Zacharov 2006: 128). Post-selection should not be done carelessly as it can remove important data, and it is also expensive in the sense that the subjects have already participated in the experiment (ibid: 128–129).

Various technical standards describe categorisations of subjects for sensory testing and listening tests (cf. Bech and Zacharov 2006: 107–112). Three categories which are included in most of these categorisations include naïve, selected, and expert. One definition of naïve subjects is described in the ITU-T P.800 recommendation (ITU 1996: 27), which describes methods for the subjective evaluation of telephone transmission quality. Naïve or untrained listeners (subjects) are defined as people who have not worked in telephone circuit assessment, have not participated in subjective tests for at

least six months and have never heard the test stimuli sentence lists before (ibid: 18). Essentially, naïve subjects are subjects who are assumed to have no expertise in assessing the samples used in the test. On the other hand, expert listeners are expected to have skills to assess the relevant properties of test samples. According to ITU-R BS.1116-2 (ITU-R 2014: 4), expert listeners can be selected based on audiometric tests and performance in previous listening tests. If the subjects are selected for a panel that is used for multiple experiments, they should be trained for the tasks the experiment requires them to perform (Bech & Zacharov 2006: 138).

Scriven (2005) argues that there are only two types of subject panels: panels for measuring properties of samples and panels for measuring the response of a wider population. Scriven (2005: 528) divides the experiencing of a stimulus into a primary and secondary response. The primary response consists of the recognition and measurement of the stimulus, and the secondary response is the formation of a judgement. Naïve assessors will typically make general judgements about stimuli, such as liking or disliking, because they are not aware of their primary response (ibid: 528). Therefore, trained subjects should be used for measuring the properties of samples and naïve subjects for measuring the population's response to the samples (ibid: 537). Using the definitions of affective and perceptual measurement (cf. Bech & Zacharov 2006: 3), naïve subjects are better for affective measurements and expert subjects are better for perceptual measurements according to Scriven's categorisation.

## 2.5. Bias and other possible issues

Getting reliable results from listening tests requires the experimenter to be aware of several types of bias effects and other issues that can cause results to be unreliable and inconsistent. Bias is defined by Lawless and Heymann (1998: 805) as "any systematic tendency to distort the overt response to a stimulus so that it becomes an inaccurate reflection of the actual sensation or hedonic reaction". Issues might result from subjects, test design and environment, or stimuli. Describing all possible types of bias is beyond the scope of this study, so only the most relevant types will be described.

It is possible that subjects give inconsistent judgements on the stimuli, leading to data which can be difficult to interpret. Especially in hedonic judgements, when subjects are asked to rate stimuli based on their personal preferences, the given answers can be very different for different subjects. This is called *between-subject inconsistency*, which can result in a multimodal distribution of results, meaning that there is not only one typical answer. If the results are multimodally distributed, calculating the mean of the results might not be at all meaningful because the mean will not represent a typical answer. This kind of results can be used, but suitable techniques, such as segmentation of subjects, have to be employed in the analysis. A subject can also change his or her mind and preferences, even during the experiment, causing the results of the subject to be inconsistent. This type of inconsistency is called *within-subject inconsistency*. (Zielinski 2006: 2.)

Subjects' expectations can affect their judgements on stimuli. A *stimulus error* can occur if the subject knows, or thinks he knows, what the sample is. This can cause the subject to have expectations which have an effect on judgement. (Lawless & Heymann 1998: 330.) Toole and Olive (1994) have demonstrated the effect of expectations in loudspeaker evaluation. Their suggestion for avoiding expectation bias is that listening tests should be conducted as blind tests if only the auditory aspects are being investigated (Toole & Olive 1994: 16). In online testing the user interface should be designed not to create expectations. This will be described in more detail in the fifth chapter of this study.

Humans typically make judgments based on a frame of reference, which can cause contextual effects and bias that affect the results (Lawless & Heymann 1998: 302). Stone et al. (2012: 169) claim that even pre-selection discrimination tests can cause contextual bias in preference tests as the selected subjects will become biased towards the differing sample in their judgements. One contextual effect is adaptation, in which the subject becomes less responsive to the stimuli as the test proceeds (Lawless & Heymann 1998: 307). Another significant contextual effect is contrast, which can cause weak stimuli to be judged weaker if strong stimuli are present and vice versa (ibid: 306).

The ordering of samples can have effects that reduce the reliability of results. A *time order error* is caused by the samples' order of presentation. Different orders of samples will cause the samples to be judged differently. Time order error can be dealt with by randomising the order or counterbalancing the order so that the samples are presented in all possible orders. (Ibid: 331.) All of the effects of context cannot be removed by randomisation and counterbalancing because the samples will always create a context for the test, and all the samples will be judged in that context (ibid: 333). Samples should be prepared in such way that they only vary in the attribute that is being measured and do not cause bias.

Scaling methods can also be a source of bias. Subjects might not be familiar with the units of magnitude they should use to make their judgements and may, as a result, make judgements that do not reflect their experiences (Bech & Zacharov 2006: 89). Subjects can have a tendency to be cautious with their use of the scale, over- or underestimating the differences in the samples, resulting in *contraction bias* (ibid: 87). Stimuli can also be rated in a way that does not reflect the differences between the stimuli (Lawless & Heymann 1998: 319), or a subject might use a scale according to his own preferences disregarding some part of the scale (ibid: 325). A well designed scale should not cause bias (Stone et al. 2012: 84).

These types of bias are very likely to occur in experiments on music perception. The judgements of musical stimuli are often hedonic (cf. third chapter of this study), which can cause different types of inconsistencies. Musical subjects can be very analytic when listening to musical stimuli and they can recognise, or think they recognise, the stimuli. This can be very evident in subjects with perfect pitch (cf. e.g. Krumhansl & Shepard 1979: 587). This can potentially cause expectation bias and stimulus errors. Musical stimuli can also create strong context effects because the stimuli can start to form a musical sequence in the subject's mind. Rating musical stimuli can potentially be difficult for non-musical subjects, leading to bias caused by scaling methods. These forms of bias must be taken into consideration when designing a listening test for investigating music perception.

# 3. OVERVIEW OF LISTENING TESTS ON MUSIC PERCEPTION

Music samples are used as stimuli in listening tests for researching various research questions in many fields of research. Music samples can be used for perceptual evaluation of audio quality, research in cognitive science, and music cognition and psychology. The perception of music is also studied using non-music stimuli. For example, sinusoidal sounds are often used in investigating the perception of pitch. This chapter describes the methods, practices and results of some relevant music related listening test experiments. Listening test experiments in three areas of research are described: musical memory and recognition, pitch perception and the perception of harmony. Musical memory and pitch perception have an important role in the perception of harmony, making those areas of research relevant for this study.

## 3.1. Testing musical memory and music recognition

Dowling and Fujitani (1971) investigated the effect of melodic contour on the recognition and remembering of melodies. Melodic contour refers to the shape of a melody. In the first experiment of their study, Dowling and Fujitani used forty-nine students from a psychology course as subjects. The subjects were not divided into separate groups according to their musical experience. In the experiment the subjects were played short standard melodies followed by a repetition of the same melody or a different melody. The subjects were divided into three groups based on the samples they were presented. For the first group the differing melody was a random selection of notes, and for the second group the differing melody had the same contour as the standard melody. The subjects in these two groups were instructed to judge whether the melodies were the same. The third group had a different task. They were instructed to judge whether the melodies had the same contour. For the third group the samples that were compared to the standard melody were either melodies with the same contour but different tones and intervals, or random melodies. The samples were synthesised using a sawtooth waveform generated with a computer. The subjects' responses were collected using a scale with four categories from "Sure Same" to "Sure Different". (Dowling &

Fujitani 1971: 526–527.) The results of this experiment indicate that transposed melodies were more difficult to identify, and random melodies were easily distinguished from the standard melodies (ibid: 528). According to Dowling and Fujitani (1971: 528), contour has an important role in the recognition of transposed melodies.

In the second experiment of their study, Dowling and Fujitani (1971: 528) tested the recognition of distorted versions of folk tunes the subjects were familiar with. In this experiment twenty-eight students were used as subjects and the samples were recordings of the tunes and their distorted versions, played on a soprano recorder. The subjects were instructed to identify the melody (ibid: 529). Some of the distorted samples retained the contour and relative interval sizes of the original melody, some retained only the contour and some retained only the first two notes. The results of this study also indicate that the contour of a melody has an important role in the recognition process (Ibid: 529–530).

Dowling (1978) has conducted further studies on the recognition and remembering of melodies. Dowling (1978: 348) used the same paradigm as in the first experiment by Dowling and Fujitani (1971) but used also tonal melodies instead of using just atonal melodies. The subjects were students, and there were twenty-one of them. In this experiment the subjects were divided into musically experienced and inexperienced subjects according to their experience. A four-category scale, just like the one in the first experiment was used. Dowling's (1978: 350) results indicate that the scale and tonal context of a melody have a noticeable impact on the recognition of melodies.

Deutsch (1972) has investigated musical memory using listening tests. The experiment in her 1972 study was designed for investigating pitch memory. The twelve subjects were played a computer generated test tone followed by six randomly selected tones and a second test tone. The task was to judge whether these two tones were the same or different. Certain tones in the sequence between the test tones were found to interfere significantly with remembering the first test tone. (Deutsch 1972: 1020–1021.)

In Cuddy and Cohen (1976), three different models of music recognition were tested by conducting a listening test experiment. The models are based on different assumptions on which intervals in a melody are used for recognition. The forty-two subjects were volunteers selected from university students (Cuddy & Cohen 1976: 258). Interestingly, the subjects were divided into three categories: untrained, trained and highly trained. Unfortunately, the criteria which the panel of judges used for categorising the subjects are not described in the study. The subjects formed three equally sized groups of fourteen subjects. A standard sequence of three notes was played, and it was followed by two transposed sequences: one identical and one different. The task was to identify which of the two sequences was the same as the standard. The stimuli were recordings of the test sequences played on a piano (Ibid: 258). Cuddy and Cohen (ibid: 259) also mention that the tests were conducted in a sound-isolated chamber. One of the important findings in the experiment was that untrained listeners are not very capable in interval recognition (ibid: 267). Regarding melodic recognition the results of the experiment do not indicate any of the three tested models as a general model of melodic recognition.

Cuddy and Lyons (1981) investigated music recognition further with two experiments. Fifty-four subjects participated in the first experiment. The participants were divided into three categories like in Cuddy and Cohen (1976). The subjects were presented seven-note sequences which varied in tonal structure (Cuddy & Lyons 1981: 18). The subjects were played a standard sequence followed by two transposed sequences. One of the transposed sequences was identical to the standard sequence, and the other one was different (ibid: 19). The test setting and task were practically the same as in Cuddy and Cohen (1976). Highly trained subjects scored highest scores and untrained subjects the lowest for recognising the correct sequence (Cuddy and Lyons 1981: 20). The second experiment of the study was not a listening test experiment and will not be considered in more detail in this study.

These experiments (Dowling & Fujitani 1971, Dowling 1978, Deutsch 1972, Cuddy & Cohen 1976, and Cuddy & Lyons 1981) have certain things in common. The subject samples are taken from student populations, the subjects are often divided into groups depending on their musical skills, and the number of subjects is not very great. The

musically skilled groups often show better skills in recognising the stimuli correctly. The method of measurement is typically a simple choice between two alternatives, except for the studies by Dowling (1978) and Dowling and Fujitani (1971), although even in their studies the responses of the subjects are dichotomised into two categories. In recognition tasks a simple "same" or "different" scale or picking the correct alternative is typically sufficient. There are also other ways to measure the recognition of musical stimuli, such as having the subjects turn towards a loudspeaker when a familiar stimulus is heard (cf. Trehub 2001).

### 3.2. Testing pitch perception

According to Attneave and Olson (1971: 1), "[t]he measurement of elementary sensations is obviously important if sensations are viewed as the building blocks of perception". It is therefore reasonable to think that pitch perception has a very important role in music perception, especially in the perception of dissonance, consonance, and harmony. A few of the most relevant listening test experiments and some new studies on pitch perception in cognitive science will be briefly described.

Stevens et al. (1937) conducted an experiment to construct a psychological scale of pitch: the *mel* scale. A subject was played two tones alternately, a single tone playing two seconds at a time. The first tone was fixed in frequency, and the second tone's frequency could be adjusted by the subject. The subject was instructed to adjust the frequency of the second tone until its frequency was half the frequency of the fixed tone. The room used for the experiment was a sound-absorbing room. There were only five subjects, and two of them were involved in designing the test. (Ibid: 186–187.) The experiment by Stevens et al. is interesting in the sense that it uses relative scaling for pitch measurement. The questionable part is the generalisability of the results as only five subjects were used. Only one of the subjects was a trained musician, and his judgements differed from those of the other four subjects (ibid: 188).

Attneave and Olson (1971: 148) consider experiments testing pitch perception with individual tones to have little relevance outside the laboratory. Instead of considering the pitch of a single tone as a perceptual object, pitch should be considered a medium, which can contain patterns. These patterns can exist in different locations of the medium, and the transposability of musical patterns supports this view. A scale for measuring pitch should also describe these properties. Two experiments were conducted by Attneave and Olson to construct a scale for measuring the perception of pitch. In the first experiment six university students with different musical abilities were instructed to adjust a response pattern to match a stimulus pattern. The stimuli were patterns consisting of two tones in various octaves, and the response pattern was always in the same octave range. The experiment was conducted in a sound attenuated-room. (Ibid: 149–150.)

In the second experiment four non-musical subjects were instructed to adjust the frequency of two tones in relation to a third tone so that the tones would create a pattern like the NBC chimes pattern used in broadcasts. (Ibid: 158–159.) The results of Attneave and Olson's first experiment indicate that the musical subjects made transpositions on a logarithmic or a musical scale, and that the responses of non-musical subjects varied greatly (ibid: 153–154). In the second experiment the non-musical subjects responded in a way that was similar to the responses of the musical subjects in the first experiment (ibid: 161). Attneave and Olson (1971: 163) conclude by stating that the mel is not a suitable unit of melody. Attneave and Olson's (1971) experiments were also conducted with very few subjects, which makes the generalisability of the results unreliable.

Cuddy (1971) conducted three experiments on the effects differently tuned intervals have on the absolute judgement of pitch. There were three sets of stimuli, and the sizes of intervals were different in each set (Cuddy 1971: 44). Seven to fifteen subjects were used in the experiments, and the task was to judge the pitch of tones by trying to recognise them. In experiment II the pauses between stimuli were also varied to investigate the effect of the previous stimulus on judgement. The judgement of some of the stimuli changed with the presentation rate (Cuddy 1971: 51). In the third experiment

the subjects were trained for the task (ibid: 51). Cuddy's experiments (1971) differed from those by Stevens et al. (1937) and Attneave and Olson (1971) in the lack of measurement scales. The perception of pitch was measured only by recognition of tones. The most important result of Cuddy's experiments was that a tone set based on a triad made the recognition of pitches more accurate (ibid: 53). Cuddy's experiments also used a relatively small sample of subjects, as did Stevens et al. (1937) and Attneave and Olson (1971), but Cuddy makes no mention of the listening environment unlike Stevens et al. and Attneave and Olson. On the other hand, Cuddy's (1971: 44) experiments were conducted using headphones, so the effects of the listening space are not that significant.

Novitski et al. (2004) have investigated the accuracy of neural and behavioural pitch discrimination. The study investigates the fundamentals of music perception. Two experiments were conducted: one using EEG (*electroencephalography*) to measure the event-related potentials during a pitch discrimination task, and the other was a behavioural experiment in which the subjects were instructed to compare the pitch of tones in pairs. During the EEG experiment the subjects were supposed to concentrate on watching a film and ignore the sound stimuli. The stimuli were either sinusoidal tones or tones with two overtones (Novitski et al. 2004: 28). The amplitude of the event-related potential was found to increase at higher frequencies and with larger pitch differences (ibid: 34), indicating that higher frequencies and larger pitch differences cause greater responses in the brain.

A study by Marques et al. (2007) investigated the detection of pitch variances in speech. The subjects were divided into musicians and non-musicians. The subjects were presented samples of speech in a language which they did not understand (Marques et al. 2007: 1453). Musicians were found to detect smaller pitch changes (ibid: 1459). The study by Marques et al. suggests that there is also neurological evidence for categorizing the subjects into musicians and non-musicians in listening tests on pitch perception.

### 3.3. Listening tests on dissonance, consonance, and harmony

Listening test experiments have been used widely for investigating the perception of consonance, dissonance, and harmony. Consonance and dissonance are often divided into two categories: sensory and musical. According to Schellenberg and Trainor (1996: 3321), sensory consonance refers to the physical properties of sound, while musical consonance is learned through exposure to music. Terhardt (1984: 282) considers musical consonance to be a combination of sensory consonance and harmony. Various theories on sensory consonance and dissonance exist, such as the theory of critical bands, which is explored by Plomp and Levelt (1965).

Levelt et al. (1966) investigated the sensory consonance of musical intervals. They conducted experiments with sinusoidal tones and tones with overtones. The tones were presented simultaneously, and the frequencies between the tones had a fixed ratio. Apart from two intervals, all the intervals were found within an octave (Levelt et al. 1966: 164). The test used a triadic comparison method: subjects were instructed to compare three intervals and select the pair that was most alike and the pair that was least alike. An incomplete balanced block design was used because otherwise all of the four subjects would have had to judge 455 different triads of intervals (ibid: 165–166). Levelt et al. (1966: 178) conclude that the method of triadic comparison was considered easy by the subjects, and that the differentiation of musical intervals was based on their width.

In their two studies on sensory consonance, Kameoka and Kuriyagawa (1969a & 1969b) attempt to form a theory of consonance. In the first study (Kameoka & Kuriyagawa 1969a: 1452) consonance is defined as "clearness" and dissonance as "turbidity". An incomplete paired comparison method with a rating scale was used in the experiments. The subjects were asked to rate the "distance in consonance" between two successive sinusoidal tones (ibid: 1452). The number of subjects varied between 11 and 35 and all of the subjects were audio engineers or students of audio engineering. A V-shaped curve was formed by the results, indicating that consonance first decreases as the interval widens and then increases as the interval starts to approach an octave

(Kameoka & Kuriyagawa 1969a: 1454). Considering the method of presenting the samples and the task the subjects were given, the measurements by Kameoka and Kuriyagawa (1969a) have very little to do with musical consonance.

In the second study by Kameoka and Kuriyagawa (1969b), experiments were conducted with a similar group of subjects and similar methods but the stimuli were tones with harmonics. The results of the second series of experiments differ from the first in the lack of a V-shaped curve. For example, the level of consonance increases with a perfect fourth and fifth (ibid: 1465).

Schellenberg and Trehub (1994a) reanalysed the results of various experiments on sensory consonance, including the experiments by Plomp and Levelt (1965), Levelt et al. (1966) and Kameoka and Kuriyagawa (1969a & 1969b). Schellenberg and Trehub (1994a: 199) came to the conclusion that ratio simplicity affects the perception of consonance, and that the experiments by Plomp and Levelt (1965) and Kameoka and Kuriyagawa (1969a & 1969b) did not test the effects of frequency ratios properly.

Schellenberg and Trehub (1994b) also conducted an experiment in another study. They tested the discrimination of an alternating pattern of sinusoidal tones of two pitches. The stimuli consisted of alternating patterns that changed at times. Forty students served as subjects, and they were instructed to raise their hand if they heard a different pattern (Schellenberg & Trehub 1994b: 474). The results indicate that changes from a pattern with an interval of a simple frequency ratio to a pattern with an interval of a complex frequency ratio were more commonly detected by the subjects. Schellenberg and Trainor (1996) conducted further experiments on sensory consonance and interval width. The experiments tested the effect of consonance on the discrimination of intervals between complex tones (Schellenberg & Trainor 1996: 3322). Sensory dissonance was found to have a greater effect on the discrimination of intervals than interval width (ibid: 3326).

McDermott et al. (2010) investigated the effects of musical experience on the perception of dissonance. The subjects were tested with both musical and non-musical stimuli

(McDermott et al. 2010: 1035). A correlation between preferring consonant chords and preferring sounds with harmonic spectra was found. It was also found that musical experience correlated with the preference of harmonic spectra, indicating that the preference for harmonic spectra is in some part learned (ibid: 1037). McDermott et al. (2010: 1037–38) conclude that enculturation has a strong effect on the perception of consonance, and that it seems that through musical experience people learn to like the acoustic property of harmonicity.

The division into sensory and musical consonance (or dissonance) is not necessarily very meaningful, especially in the context of music psychology and musicology. In experiments on sensory consonance, subjects would have to somehow completely ignore their musical backgrounds and enculturation, which seems an impossible task. Many of the experiments on sensory consonance have used sinusoidal tones, which do not reflect real-world auditory experiences, as stimuli, making it difficult to generalise the results to realistic scenarios. The study by McDermott et al. (2010) also suggests that the perception of consonance even in non-musical stimuli is affected by musical background and enculturation, thus making the notion of a purely physical phenomenon of consonance questionable. Considering also the findings of neurological research on pitch perception (cf. e.g. Marques et al. 2007), there is evidence that musical experience affects pitch perception fundamentally.

Carol Krumhansl, among others, has conducted numerous experiments on the perception of musical consonance and tonality. Considering the fact that the purpose of this thesis is to design a listening test experiment for investigating the perception of musical consonance, the studies by Krumhansl and others offer very relevant insight into the design of such an experiment.

A listening test experiment for determining the perceived similarity of tones in different contexts is described by Krumhansl (1979: 350). The subjects were instructed to judge the similarity of two succeeding tones that were preceded by a major triad, an ascending major scale or a descending major scale (Krumhansl 1979: 353). According to Krumhansl (1979: 358), the results of the experiment indicate that, in a musical context,

musically experienced subjects perceive the relationships between the individual tones and judge diatonic tones as more closely related than non-diatonic tones. Interval width also had an impact on ratings of similarity: tones separated by a small frequency difference were rated more similar than tones separated by a large frequency difference (ibid: 359). The study also includes a second listening test experiment on the effects of diatonic and non-diatonic sequences on pitch memory, but this experiment is not relevant for the purposes of this thesis.

In the study by Krumhansl and Shepard (1979), two very similar listening test experiments are described. The subjects were instructed to judge how well a tonal context was completed by a succeeding tone. In the first experiment the test tones were selected from the chromatic scale (Krumhansl & Shepard 1979: 583), and in the second experiment the tones were selected from a scale that also included quarter tones (ibid: 589). The results of both experiments indicate that pitches are judged differently in a musical context than in the experiments on sensory consonance, and that certain intervals such as the perfect fifth and major fourth are judged considerably more consonant than other intervals close in frequency (ibid: 586–587 & 592). A very interesting result was obtained in the first experiment (ibid: 587): a subject who was reported to have absolute pitch rated the tones of a major triad using the highest mark on the rating scale, which would indicate that the decision was probably affected by the recognition of the tones and knowledge of music theory.

The studies by Krumhansl and Kessler (1982) and Krumhansl and Keil (1982) utilised experiments in which subjects were yet again instructed to judge how well tones completed different contexts. In the first experiment by Krumhansl and Kessler (1982: 341), the subjects were played various different scales, chords and cadences to create a tonal context which was followed by a single tone. The results show a preference for the tones of the tonic triad (ibid: 343). Krumhansl and Keil (1982) conducted an experiment also including children as subjects. The subjects were asked to help the experimenter write a song and were instructed to rate the sequences they were played (Krumhansl & Keil 1982: 246). The rating scale for the children used a frowning and a smiling face for anchors. The stimuli were sequences that began with the four tones forming a major

triad and were followed by two tones that were either diatonic or non-diatonic (ibid: 246). Diatonic tones were yet again considered to fit better into the context than non-diatonic tones, even by young children (ibid: 249).

The study by Krumhansl et al. (1982) differs from the earlier studies in which Krumhansl has been involved. Instead of single tones chords were rated. The stimuli were sequences of two chords which were preceded by an ascending scale. The subjects were instructed to judge how well the second chord followed the first (Krumhansl et al. 1982: 28). The scale context was found to have a relatively weak effect on the judgements, which might have been caused by the fact that chords are musically more complex than single tones or by the fact that three different scale contexts were used in the experiment (ibid: 33).

Bharucha and Krumhansl (1983) conducted further experiments on the perception of chords and tonal context. In experiment 1 a cadence followed by two chords was played to the subjects, and the subjects were instructed to judge how well the second chord followed the first. A no-context condition, in which the two chords were not preceded by a cadence, was also tested (Bharucha & Krumhansl 1983: 75). Chords that are from closely related but different keys were found to be judged less similar than chords from the same key (ibid: 82).

Other studies on the perception of harmony and tonality have been conducted by Janata et al. (2003), Brattico et al. (2008) and Virtala (2015). In the experiment by Janata et al., the subjects tried to detect wrong or out-of-place notes from music samples. Brattico et al. investigated the neural response to chords that do not fit into the musical context they appear in. Virtala also conducted similar experiments for her dissertation. Both Brattico et al. (2008: 2241) and Virtala (2015: 69) found musicians to have an enhanced neural discrimination of the out-of-place chords.

All of the aforementioned studies on the perception of tonality that involve Krumhansl have many things in common. First of all, the stimuli attempt to form a musical context. The subjects rate the stimuli using a seven-point interval scale and the results are

analysed using multidimensional scaling (cf. Kruskal 1964). Multidimensional scaling is used to create geometric representations of pitch, which Krumhansl (1990: 112) considers suitable for representing the dependencies between the pitches. The task the subjects are presented is to judge either the similarity of stimuli or how well the stimuli fit into a context. On average the subjects also tend to have many years of musical experience in all of the experiments. The subjects are mostly university students, and the sample size is not typically very large.

Although the term *musical consonance* is not used in Krumhansl and Kessler's (1982) study, the first experiment essentially tests the musical consonance of tones in relation to a tonal context. That experiment is very important in the context of this thesis as the experiment designed in this thesis aims to test how well subjects perceive tones to fit into the context of a power chord.

# 4. INTERNET EXPERIMENTS AND ONLINE LISTENING TESTS

Internet technology has developed rapidly, and high-speed connections have become more common during the 21st century (Kurose & Ross 2013: 91). According to websites such as internetlivestats.com (Internet live stats 2015) and internetworldstats.com (Internet World Stats 2015), there are over three billion Internet users in the world today.

According to Reips (2002b: 244), if an experiment is conducted in a laboratory using a computer, the experiment can also be conducted as an Internet experiment without losing anything. This is not completely the case with listening tests as the experimenter loses control over background noise level and the audio equipment connected to the computer. Internet experimenting does have various advantages, which can make up for the loss of some control. Reips (2002b: 244) lists many advantages of Internet experiments including speed, low cost and a wider sample of subjects.

## 4.1. Sampling from the Internet population

Sampling from the Internet can be done using various methods. Hewson et al. (2003: 36–42) describe different methods of sampling from the Internet population. Samples can be either volunteer or non-volunteer samples. Volunteer samples are typically obtained by posting an announcement about the study and asking people to participate. Volunteer sampling is also called self-selection (cf. Reips 2002b: 247). Non-volunteer samples are obtained by pre-selecting the subjects and asking them personally to take part in the experiment. Non-volunteer samples are quite problematic to obtain from the Internet because it would require the experimenter to have e-mail addresses for all the potential subjects, and it is not possible to estimate non-response bias without knowing more about the subjects than the e-mail address tells. Also volunteer samples are considered problematic because they can be biased. The people who are most likely to participate in the survey or experiment voluntarily do not necessarily reflect the population. With volunteer sampling it is also hard to have a good understanding of the

sampling frame, that is, all of the people who saw the announcement and could have participated in the experiment. Despite these problems, volunteer sampling is the most common method of sampling from the Internet. (Hewson et al. 2003: 36–42)

Self-selection is not considered to cause serious problems in experiments that investigate perception or other topics in which people are not thought to vary significantly (Reips 2002b: 247). In music perception people can vary a lot in their musical skills, which affects their responses. Andrews et al. (2003: 193) point out that the skills and abilities of potential subjects affect their decision whether to participate in an experiment or not. It can therefore be assumed that it is more difficult to recruit non-musical subjects when it comes to conducting a music perception experiment online. According to Balch (2010: 81), an announcement about an experiment on any website can only get subjects who are interested in both the website and the experiment. Therefore, using music-related websites to recruit subjects for music-related experiments will probably attract subjects that are interested in both music and the experiment. On the other hand, recruiting non-musical subjects for an online music perception experiment can be difficult.

Even when a subject has shown enough interest in the experiment to actually begin the experiment, the subject can still abandon the experiment before completing it. This is called *dropout* (Reips 2002b: 248). Many reasons can cause subjects to quit the experiment. If the experiment takes too long, the subject can get bored and go to more interesting websites (Hewson et al. 2003: 83). According to Bowers (1998: 47), 20 minutes is a suitable time for an online survey, which is fairly similar to the time recommendation that Bech and Zacharov (2006: 302) make for laboratory-based listening tests. Considering the changes the Internet and the culture related to it have gone through during the last decade, it is not certain that Bowers' suggestions are relevant anymore. Simsek and Veiga (2001: 224) point out that measurement errors can occur in the form of dishonest answers if subjects are asked questions they do not want to answer. It is very possible that subjects can quit the experiment for the same reason. There are some simple ways to reduce dropout. For example, Reips (2002a: 243) suggests placing personal information questions in the beginning of the experiment

because this way people are less likely to quit the experiment before completing it. Reips (2002a: 242) also suggests collecting data on dropout rate and using it as a dependent variable.

One of the big questions concerning research conducted on the Internet is whether the sample selected from the Internet population is representative of the real population of people. According to Hewson et al. (2003: 27), it is often claimed that "the Internet-user population constitutes a dramatically skewed sample of the 'population at large' – ". Hewson et al. (2003: 28) note that various studies from the late 1990's on the subject of Internet sampling are not conclusive on Internet samples being any less representative of the whole population than traditionally selected samples. Hewson et al. (2003: 29) also remind of the fact that Internet samples that resemble traditionally selected samples are not necessarily good samples, because traditional sampling techniques are not perfect either. Considering that the text by Hewson et al. is from 2003, and that the population of Internet users has grown and diversified dramatically, the concerns over the Internet population being skewed is no longer as relevant.

Many of the listening test experiments on music perception described earlier in this study have used samples of mostly university students, and the sample sizes have rarely exceeded 50. Using online listening tests, it is possible to gain much larger and more diverse samples, potentially making the results more applicable to a wider population. One of the possible challenges in online listening tests on music perception can be recruiting non-musical subjects using volunteer sampling.

### 4.2. Reliability of subjects and data

Internet surveys and experiments have many potential problems regarding the reliability of data. Subjects may give dishonest answers, or they can take part in an experiment multiple times. Subjects can participate in an experiment in a noisy and distracting environment, which can affect the reliability of results. Internet surveys and experiments are self-administered, and as a result, the layout and graphical interface of the

experiment can cause bias (cf. Christian & Dillman 2004).

McGraw et al. (2000) investigated the reliability of Internet experiments on reaction time and recognition by analysing data from experiments conducted over a period of two years. The focus of the investigation was on the effects of the test environment. The subjects took part in the experiments using different computers with different operating systems in various locations. Different environments and settings were found to have no remarkable effect on the test results, and the possibility of obtaining a large sample of subjects outweighed the lack of control over test environments (ibid: 502). McGraw et al. (2000: 506) conclude that if the experiment is conducted in a laboratory with a computer, it can also be implemented as an Internet experiment although they do question the possibility of conducting Internet experiments which require extremely accurate measurements of reaction time (ibid: 505).

Smith and Leigh (1997: 499–500) point out that if it is possible for the subjects to participate in the experiment multiple times, they may do so and invalidate the data. Balch (2010: 90) suggests tracking IP addresses as one possible way of detecting multiple entries but also recognises the weakness of the method. An IP address is not in any way tied to a person or even a computer. A computer typically obtains an IP address for its Internet connection by using *Dynamic Host Configuration Protocol* (DHCP) from a DHCP server. The server has a collection of addresses it can allocate, and a computer is not necessarily always given the same address even with the same connection (Kurose & Ross 2013: 371–372). It is also possible that a computer is connected to the Internet using a NAT-enabled router. Every computer in a subnet that is connected to the internet via a NAT-enabled router uses the router's IP address (ibid: 376). Simply blocking an IP address when it has been used once is not sufficient because it does not stop a person from participating in the experiment again, and blocking an address can very well stop someone from participating even for the first time. Giving participants user identifications and passwords could be used effectively to block multiple entries, but it could reduce the number of participants due to the added complexity (cf. Balch 2010: 89).

There is also the possibility of the subjects giving dishonest answers, especially if the subjects may receive a price by participating (cf. Im & Chee 2011: 383). Fortson et al. (2006) compared the data from Internet and traditional surveys on traumatic stress. Fortson et al. (2006: 718) conclude that the data gathered from university students with the Internet survey was reliable. Sampling from a different population could produce different results, and subjects also might have different incentives for giving false information in different types of experiments. In a listening test experiment, if a price is not promised, there is little to be gained from dishonest answers, especially if the subjects are anonymous.

Using complicated technologies in an online experiment can cause people not to participate (cf. Reips 2002b: 248). The point of avoiding rarely used plug-ins and other technologies is relevant, yet the technologies Reips lists are not that rare anymore. For example, JavaScript is very commonly used today: according to W3Techs (2015), almost ninety percent of websites use JavaScript. Another technical consideration is security. The results should be saved in a way that they cannot be accessed by anyone who should not have access to them. Security is a very important issue when personal data is not collected anonymously, and even when the data is anonymous, it should be protected from hackers who may wish to do harm. The issue of internet security is a complicated technical issue, and it is beyond the scope of this study. Nonetheless, Internet security cannot be disregarded in Internet experimenting.

### 4.3. Music-related online listening tests

One of the advantages of online testing is that subjects can participate in an experiment in a familiar setting (Reips 2002b: 247). This can have a very notable effect on the validity of online listening tests on music perception because it is even possible for subjects to listen to the stimuli using the same equipment they normally use for music listening. The lack of control over the test equipment does not necessarily have a significant effect on the subjects' judgements (Disley et al. 2006: 66). The large number of subjects obtainable with online testing can average out the differences in test

environments and listening equipment. It is possible to obtain very large samples with online tests. For example, Cox (2007: 2) managed to recruit around 130,000 subjects for his study on the most horrible sounds in the world. Cox's experiment was publicised in many ways that might not be available for all experiments. Newspapers, television, radio and other media were used to spread the word (ibid).

Psychoacoustic research typically requires strict control over the test setting, environment and equipment. Despite this, Disley et al. (2006) have used online listening tests for investigating the connection of various verbal labels and timbre. Before the actual test a pilot experiment was conducted with sixteen musically trained subjects (ibid: 62). The pilot experiment was found to be very important as it provided valuable information on the samples. The subjects in the actual experiment were all students or staff of music and music technology departments, and the technical requirements for participating were a soundcard and quality headphones (ibid: 64). An experiment was conducted also in controlled settings, and it was found that the results from the uncontrolled and controlled groups did not differ significantly (ibid: 66). If online listening tests can be used successfully for psychoacoustic testing, which possibly requires more controlled experimental settings, there is no obvious reason why online listening tests could not be used for music perception experiments just as well. It is still important to consider the fact that in the experiment by Disley et al., the subjects probably had a good knowledge of audio technology and understood the effects of background noise, so they possibly performed the tests in reasonable conditions with good quality equipment.

Online listening tests have been used for investigating the emotional responses to music. Egermann et al. (2006) used a Java Applet for investigating emotional responses to different musical stimuli. The experiment by Egermann et al. (2006: 180) was very long, taking around 45 to 60 minutes. The subjects were recruited using e-mail lists and personal invitations, and the subjects were assigned user accounts and passwords (ibid: 179). 87 of the 107 invited subjects completed the experiment, which is surprisingly many considering the length of the experiment. Had the subjects been volunteers who had simply come across a link to the experiment on a website, more subjects would

have probably dropped out of the experiment before completing in. Also, most of the participants were musically trained to some extent (ibid: 181) and might have been more interested in such an experiment than the average Internet user. Egermann et al. (2006: 182) conclude that online listening tests seem to be suitable for studying music and emotions.

Similar experiments have also been conducted by Kosta et al. (2013) and Song et al. (2013), but they used a simple website instead of a Java Applet. These experiments also used a two dimensional measurement consisting of valence and arousal, like the experiment by Egermann et al. (2006). The experiments by Kosta et al. (2013) and Song et al. (2013) are not very relevant in the context of this study, except for the way musical expertise has been measured in these experiments. Both of the studies used questions from the Goldsmiths Musical Sophistication Index (cf. Goldsmiths 2015) for categorising subjects into different groups based on their musical experience. The subjects were asked to evaluate their musical skills by giving ratings of agreement to statements such as "I can't read a musical score" and answering questions like "How many musical instruments can I play?" (cf. Kosta et al. 2013: 319). These answers were used to calculate an overall rating of musical skills. The problem with such questions is that people can have very different views on what it means to be able to play an instrument or to read a score. Some people might think that elementary music reading skills mean they can read a score, and some might consider score reading skills to include sophisticated skills in sight singing. The same applies to many of the other questions. Asking subjects simple and specific questions, such as “how many hours do you practice each week?” and “for how many years have you taken lessons?” would probably give less biased and inconsistent results about musical background.

Online listening tests have been used to investigate the perception of rhythm in music. Wright (2008) created a downloadable application for measuring the perceptual attack time of different stimuli. The 57 subjects were recruited by sending email to selected musicians and computer music researchers (ibid: 71). In one part of the experiment, the subjects were instructed to align sounds so that their perceived attacks would occur at the same time (ibid: 77). After completing the experiment the subjects sent the results to

the experimenter via email (ibid: 72). Some of the results were discarded because the subject had accepted a random initial ordering of stimuli. The results of one subject from the pilot test were discarded as well because the subject had used poor quality speakers in a noisy environment (ibid: 78). There was no need to check for multiple entries from subjects because they were personally recruited. At the end of the experiment, subjects were asked what kind of speakers or headphones they used for the experiment (ibid: 72).

Honing (2006, 2007), and Honing and Ladinig (2006a, 2009) have investigated the relationship of tempo and rhythmic timing in music. All of these experiments were conducted online using a fairly simple website for collecting data, and in all of these experiments the task was to recognise which musical samples were tempo-modified. The recruitment of the subjects was done by e-mail and posts on Internet forums, and it was quite successful as all of the experiments had more than a hundred subjects. The experiment by Honing and Ladinig (2009) is perhaps the most sophisticated experiment of the experiments by Honing, and Honing and Ladinig. In this online listening test experiment, the subjects listened to musical samples in different genres (jazz, rock, and classical) and were instructed to decide which of the samples were tempo-modified (ibid: 284). The subjects were divided into three groups based on their musical skills, and they were also asked which musical genre they listened to mostly (ibid: 282). The results indicate that exposure to a certain type of music makes the listener better at judging the rhythmic nuances of musical samples of that genre (ibid: 287). Bigand and Poulin-Charronnat (2006) demonstrate that exposure to music has similar effects as musical training on the perception of harmony. When categorising subjects in music perception listening tests, it is important to ask what kind of music the subjects listen to typically as it has an effect on their judgements.

There are some good practices that can be learned from the online listening test experiments described in this chapter. Before the actual experiment begins, the subject should be asked the necessary information about musical training and listening habits. Setting the volume to a comfortable level should also be done before the test begins (cf. Honing & Ladinig 2009: 284). The questions on musical background should be

unambiguous and clear, leaving little room for different interpretations. With e-mail invitations to pre-selected people it is more difficult to obtain large samples, but it is possible to use a more complicated downloadable listening test application if one is necessary. Honing and Ladinig (2008: 6) consider online listening tests to be no more problematic than laboratory tests, and what is lost in internal validity is gained in external validity.

Most of the methodology described in the second chapter of this study applies to online listening tests as well. Tests must be carefully planned, and their objectives need to be clearly defined. Problems with uncontrolled environments can be reduced by asking the subjects to use headphones because their use reduces the effects of the environment (Bech & Zacharov 2006: 228). Pilot experimenting is just as important for online listening tests, and the pilot experiment should also be conducted online to make it more like the actual experiment being designed.

Overall, Internet experiments seem to be a promising method for investigating the perception of music. The potential problems with sampling from the Internet population are mostly the same problems that one faces when recruiting subjects for a traditional laboratory-based listening test. Subjects have been shown to be practically as honest in online tests as in traditional tests, making concerns over reliability no more significant than in laboratory-based tests. There are still challenges, such as avoiding multiple entries and reducing dropout. Often a trade-off must be made. For example, using strict control over participation to avoid multiple entries can cause dropout. Avoiding multiple entries can be next to impossible because reliable identification of subjects online is typically beyond the technology available to researchers. Making participation as easy as possible will result in a large sample which can cancel out the problems caused by some subjects submitting their results multiple times. In addition, collecting data anonymously reduces the damage an Internet security attack can cause, because no-one's personal information is collected.

A large sample of Internet users is likely to be representative of a wider population, especially nowadays that the Internet population has become very diverse. This makes online listening tests perhaps more valid for investigating the perception of stimuli than for investigating their properties. As music perception experiments are more concerned with the subjects' perception, online listening tests are a very suitable method for such experiments.

# 5. THE PILOT EXPERIMENT

A pilot experiment was conducted to test three different experimental methods. The goal was to find an experimental arrangement for the online listening test that would be easy for the subjects, would not take too long, and would not be boring or frustrating. The pilot experiment was also conducted in order to find any potential problems with the audio samples, the questions on musical background and the task presented to the subjects. The main research question of the pilot experiment was to find out which of the three methods is the most suitable for a large scale online listening test on the perception of heavy metal harmony.

## 5.1. Methodology of the pilot experiment

The pilot experiment consisted of five blocks: two questionnaires and three listening tests. The different blocks of the experiments were structured as follows: a questionnaire on musical background, listening test 1, listening test 2, listening test 3, and a general questionnaire on the listening tests. Each of the listening tests also included a small questionnaire concerning that test, filled in at the end.

The pilot experiment was conducted as an Internet experiment so that it would be more like the actual experiment that was being designed. This way the subjects could complete the tests at any time they wanted to, and they could complete the tests at home, using their own audio equipment. Without an experimenter present, the subjects had to rely on the written instructions, hopefully revealing any problems with the instructions during the pilot experiment. Since an Internet experiment makes it easier for the subjects to drop out, it also made possible the gathering of data on which method was likely to cause the greatest dropout rates in the actual experiment. The subjects were given one to two weeks to complete the experiment, depending on the time at which they were recruited. It was not considered likely that subjects would drop out of this experiment, because the subjects were personally recruited and showed interest in the experiment.

The experiment was implemented without any server-side scripts or programs, so the subjects downloaded a set of files containing the experiment onto their computers. The questionnaires and listening tests were implemented as html-documents that opened in the subjects' Internet browsers just as a website would open. The audio samples were contained in the same folder the subjects downloaded so that they could be embedded into the html-documents. The sending of the results was handled using html's mailto-method (cf. Duerst et al. 2010), which opens an e-mail message, with the results filled in the message body, in the subject's e-mail client. This same method of sending results was used by Wright (2008: 166) in his experiment. All of the tests started with a calibration sound that the subjects could use to adjust their volume to a comfortable level before the actual test started. This method has been used in various online listening tests (cf. e.g. Honing & Ladinig 2009: 284).

During the questionnaire on musical background and the listening tests, a timer was ran in the background in order to gather data on how long the tests took. In the listening tests the timer was stopped before the short questionnaire at the end of the test so that the timer would only measure the time taken to complete the listening part of the test. The audio samples were presented in random order. The order was not randomised separately for each subject, but instead each subject was presented the samples in the same random order. Randomising the samples separately for each subject will be necessary in the actual test, but it was not considered that important for the pilot experiment as it was conducted to gather data on different experimental methods, not on the perception of the intervals.

The questionnaire on musical background was based on the questionnaire that Honing and Ladinig (2006b) used in their experiment on the effects of exposure on the perception of timing in music. Some of the questions were omitted, such as the question on recognising various genres, while some new questions were added. The subjects were asked which instruments they had played to find out if any of them had played electric guitar because having played electric guitar can have an effect on the way a person perceives chords played with distortion (cf. Lilja 2009: 114 & 150–151).

In the first listening test, the subjects were instructed to rate how well the tones in the samples fit together musically. A scale from one to seven was used: a rating of one meant that the tones fit together poorly and a rating of seven meant that the tones fit together well. Each of the twelve samples was used twice during the test, so there were 24 trials in total. Rating the samples with a scale from one to seven was selected as one of the methods because this method has been used in numerous experiments involving Krumhansl (cf. chapter three of this study). It was considered an especially relevant method because Krumhansl and Kessler (1982) have used it for the task of judging how well a tone fits into a context.

The second listening test used the method of paired comparison. With twelve samples there are 66 different possible pairs (cf. Thurstone 1927b: 379). The 66 trials were divided onto four pages in the listening test to reduce the amount of scrolling down. Each trial had one sample marked ‘A’ and one marked ‘B’, and the subjects were instructed to select the one in which the tones fit together better musically. The paired comparison method was selected because it is a widely accepted method in psychometric tests, and the task of comparing pairs is considered an easy task for the subjects (Hynninen 2001: 10).

The third listening test used rank-order paradigm (cf. Hynninen 2001: 12) as its method. The subjects were instructed to rank the samples from best to worst based on how well the tones in the samples fit together musically. The best fit was to be ranked as 1 and the worst fit as 12. Rank-order paradigm was selected as the third method to be tested in the pilot experiment because it is widely considered to be a simple and easy way to place samples on an ordinal scale (ibid).

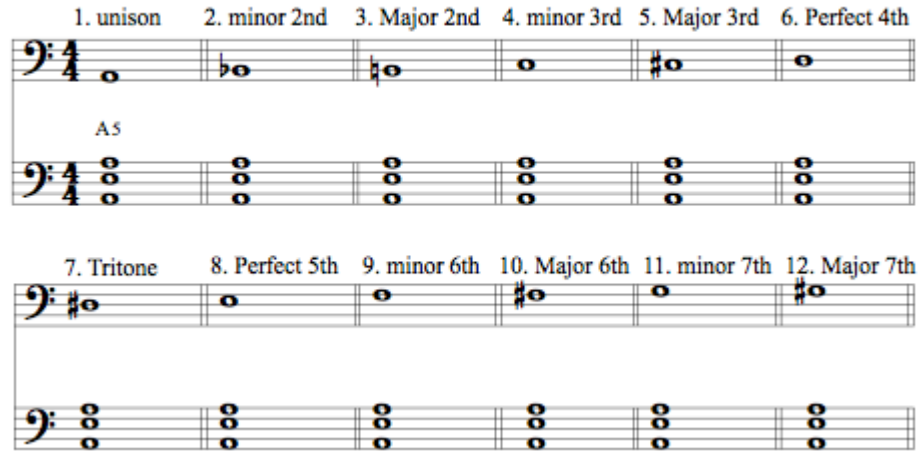
The last block of the pilot experiment was a questionnaire on the whole pilot experiment. The questionnaire contained questions on the listening tests and the questionnaire on musical background. The goal of this questionnaire was to find out which listening test the subjects had liked the most. The subjects were also asked for general comments on the pilot experiment and its listening tests. The pilot experiment's questionnaires and sample screenshots of the listening tests' html-documents can be

found in appendix A1.

In total, thirteen subjects were recruited for the pilot experiment. The subjects were sent personal invitations through social media, or they were asked to participate in person. Nine of the subjects were students of musicology, and four of the subjects were not musicologists or professional musicians. The goal was to recruit subjects who would be highly likely to complete the experiment, which is why so many of those who showed interest in the experiment and the research subject were, in fact, musicologists. Non-musicologists were recruited in order to gain a different perspective. The subjects were given a deadline of one to two weeks to complete the experiment, and those subjects who had not started the experiment during the first week were reminded of the experiment a week before the deadline.

The samples consisted of an A power chord and a single simultaneous tone within the same octave as the power chord, all played on an electric guitar. Example 1 shows the twelve samples in standard notation. Using a Squier Stratocaster electric guitar with the bridge pick-up on, the samples were recorded into Logic 8 digital audio workstation through a Radial J48 DI-box and Fireface 800 line-input. No effects were used on the signal so that the same signals could be re-used with different effects. The samples were distorted afterwards using Logic's Guitar Amp Pro plugin, which models a guitar distortion effect digitally. The same settings were used for all of the samples. The levels of all samples were adjusted to be just below clipping and to be of equal volume. The power chord and the tones were recorded separately, and they were mixed together afterwards. A fadeout curve was applied to all the samples at around four seconds so that all of the samples would be of the same length and would fade out the same way. The samples were recorded by musicology student Tommi Kotilainen and Dr. Esa Lilja, after which I mixed the power chords and tones together.

**EXAMPLE 1**




The compressed folder that the subjects downloaded contained a file with the general instructions for the entire pilot experiment (see appendix A1.1.), and all of the listening tests contained their own specific instructions. The subjects were instructed to fill in the questionnaire on musical background first, then complete the listening tests from test 1 to test 3 and last to fill in the general questionnaire on the experiment. Taking breaks between the tests was recommended, and it was noted that the subjects could complete different tests on different days. The purpose of the tests was to measure the perceived musical consonance yet the subjects were not asked to rate the samples based on their consonance as the term is probably not meaningful to non-musical subjects. Instead, in all of the three listening tests, the subjects were instructed to judge the samples based on how well the tones in the sample fit together musically. The purpose behind this was to phrase the task in a way that would be understandable to both musicians and non-musicians. This procedure was based on the one by Krumhansl and Kessler (1982: 342), in which the subjects were instructed to rate the stimuli tones based on how well they fit into a musical element. The instructions in the pilot listening tests included the word *musically* to emphasise that the judgement should be based on considering the samples as musical elements instead of judging the samples based on other auditory properties like their harshness or sound quality.

## 5.2. The results of the pilot experiment

Out of the thirteen subjects who expressed interest in participating in the experiment, eight completed it. Only one of the dropouts informed me before the deadline that he could not find the time to complete the experiment. The other dropouts were asked for a reason why they did not complete the test after the deadline. They reported that they had either forgotten about the experiment or they had not found time to complete it. Before any of the subjects had started the test, it became evident that using an e-mail client program is not very common, and most of the subjects first had to configure an e-mail client before they could start the experiment. In the case of one subject, pressing 'submit' did not even open an e-mail client on his computer, so he copied the questions from the html-document into a text file and wrote his results into that instead. Once he was ready, he sent the text file to me as an e-mail attachment. Using html's mailto-method for sending results is therefore not a very good method for collecting data, even with such a small sample.

Using a timer to measure how long it took to fill in the questionnaire on musical background and to complete the listening tests worked fairly well although there were a few cases in which a subject had opened the html-document before actually starting the test. The time data for the musical background questionnaire of two subjects and for listening test 2 of one subject were discarded because they had opened the document much before starting the test. In addition, there is no time data for the subject whose computer did not open an e-mail client on pressing the 'submit' button. Online tests always contain the risk that a subject opens the test in another window beforehand and visits other websites before actually starting the test. Therefore, the timer in an online test should not be started when the document is loaded but instead when the actual test is started, for example, by pressing a button.

One more problem occurred during the pilot experiment. One subject ranked the samples in test 3 in opposite order. He ranked the sample with the best fitting tones as 12 and the sample with worst fitting tones as 1. After I made sure that he had used the ranking this way, I reversed his ranking to make it comparable with the rest of the

ranking data.

All of the subjects had had instrument or vocal training and had studied music theory and ear training to some extent. The subjects were mostly enthusiastic music listeners as the average rating for listener type was over 4 although their music listening habits varied greatly. Only one of the subjects had perfect pitch. Filling in the questionnaire took under three minutes on average. The results of the musical background questionnaire are found in appendix A2.1.

With such a small sample, the statistics of the musical background questionnaire are not very interesting and are thus not analysed in detail. The most interesting part of the results is the comments on the musical background questionnaire. One subject noted (see appendix A2.1. comment 1) that having certain genres as categories in the questionnaire can steer the answers to questions. This can, of course, lead to bias in these answers. The reason for having categories is to simplify the analysis of music listening habits: it is easier to calculate the distribution of genres when there is a limited number of known categories. If the subjects simply described their music listening habits in free text, analysing the results would require reading through all the descriptions and parsing the data before the results could be properly analysed. In the same comment this subject noted that he had placed most of his music listening in the 'other' category. Having the 'other' category can therefore be considered important. The approach of having some general categories ready on the questionnaire can simplify the analysis of results, and having an 'other' category enables the subjects to describe their listening habits in more detail if they wish to. Despite the subject's comment, changing the questions on music listening habits is not necessary as they do not have a major role in categorising the subjects.

Other important things that the subjects pointed out about how the musical background questionnaire could be developed were that they could be asked to specify the type of musical training (see appendix A2.1. comment 4) and that music theory training and ear training could be asked separately because they are not always taught or studied together (see ibid: comment 3). Making these changes could help to categorise the

subjects more accurately, and only two more questions would be required. Therefore, these questions were added to the musical background questionnaire for the design of the large scale experiment.

The results of the pilot experiment's listening tests will only be described briefly because they are only of secondary importance, and their results cannot be considered generalisable due to the small sample size and the lack of randomisation in the tests. On the other hand, the results of the tests can be compared to each other to see what kind of effects the test method can have on the judgement of samples. The results of the listening tests are found in appendix A2 (see A2.2.–4.).

The data from test 1 was used to calculate an average rating for each sample and these ratings were used to place the samples on an ordinal scale. The data from test 2 was used to calculate, for each sample, the number of times the sample was preferred to other samples. This way the total number of preferences could be calculated for each sample, and these numbers could be used to place the samples on an ordinal scale. This method of handling the data was based on the method by Kendall and Babington Smith (1940: 343). The data from test 3 was handled similarly to those of test 2. The numbers indicating rank were mapped into the number of times the sample was preferred to other samples by subtracting the rank number from twelve. These numbers were used to calculate the total number of preferences for each sample. Using the total numbers of preferences the samples were put in order. The ordinal scale results of the three tests are presented in Table 1.

**Table 1**

**Test 1**
*Most consonant*

| | |
|---|---|
| 1. | unison |
| 2. | Perfect 5th |
| 3. | Major 3rd |
| 4. | minor 7th |
| 5. | Major 2nd |
| 6. | Perfect 4th |
| 7. | Major 6th |
| 8. | minor 3rd |
| 9. | Major 7th |
| 10. | Tritone |
| 11. | minor 6th |
| 12. | minor 2nd |

*Least consonant*

**Test 2**
*Most consonant*

| | |
|---|---|
| 1. | unison |
| 2. | Perfect 5th |
| 3. | Major 3rd |
| 4. | Major 2nd |
| 5. | minor 7th |
| 6. | Major 6th |
| 7. | Perfect 4th |
| 8. | minor 3rd |
| 9. | Major 7th |
| 10. | minor 6th |
| 11. | Tritone |
| 12. | minor 2nd |

*Least consonant*

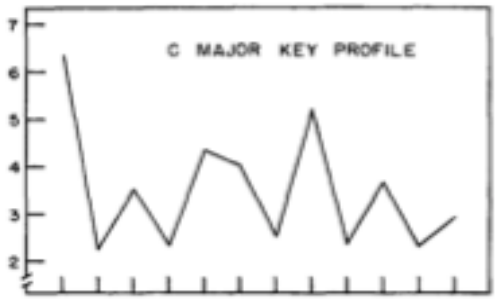
**Test 3**
*Most consonant*

| | | |
|---|---|---|
| 1. | unison | |
| 2. | Perfect 5th | |
| 3. | Major 3rd | |
| 4. | minor 7th | *4. and* |
| 5. | Perfect 4th | *5. are tied* |
| 6. | Major 2nd | |
| 7. | minor 3rd | |
| 8. | Major 6th | |
| 9. | Major 7th | |
| 10. | Tritone | |
| 11. | minor 2nd | |
| 12. | minor 6th | |

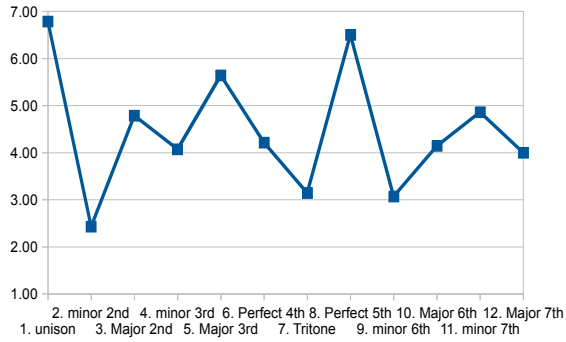
*Least consonant*

The results of each test were used to create a profile similar to the figures by Krumhansl and Kessler (1982: 343). Figure 1 below contains graphs of the results from Krumhansl and Kessler's (ibid) experiment using a major scale as a context and the results of the three tests of the pilot experiment. There are many similarities between Krumhansl and Kessler's major context experiment's results and the results of the tests in the pilot experiment. Unison, perfect fifth and major third are ranked highest while tritone, minor sixth and minor second are ranked very low. The most notable difference between the results of the pilot experiment and Krumhansl and Kessler's experiment using a major scale as a reference is that the sevenths are ranked differently. In the musical context of the power chord the minor seventh is ranked more fitting than the major seventh, while in Krumhansl and Kessler's experiment the major seventh was ranked more fitting than the minor seventh. With such a small sample, these results cannot be used to make any general conclusions about the consonance of intervals in relation to a power chord. Despite the small differences in the ordinal scales, the similarity of results from the three tests makes it evident that the method of testing does not have any significant effects on the results. Therefore, any of these methods should be valid.

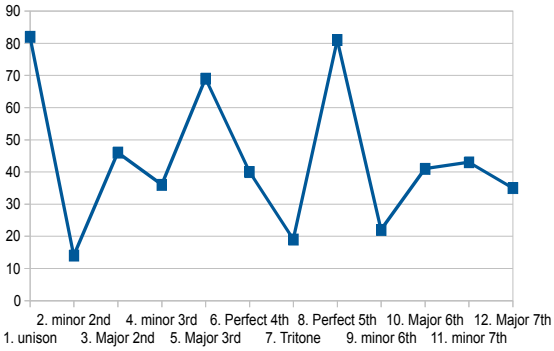
Figure 1



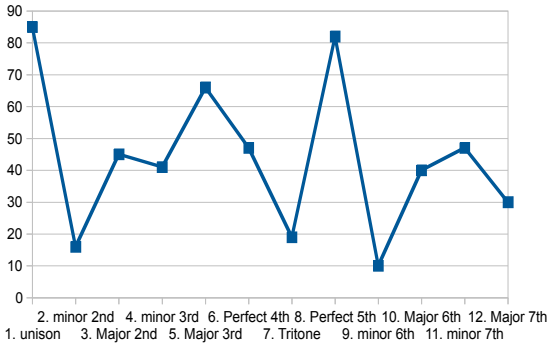

*a) from Figure 1, Krumhansl & Kessler (1982: 343)*




*b) results from test 1 of the pilot experiment*




*c) results from test 2 of the pilot experiment*




*d) results from test 3 of the pilot experiment*

Based on the comments test 1 was found to be monotonous. As test 1 was the first the subjects completed, many of the comments on the meaningfulness of judging how well tones fit together musically were found in test 1's comments. Also the meaningfulness of the scale from 1 to 7 was questioned. One subject noted that only after listening to many samples, he started to think how he should judge the samples, and another noted that he thought that using the scale required listening through all of the samples to use the scale in a logical way. The higher end of the scale was mostly used by one subject who said he liked "weird intervals". He ranked all of the samples quite high (A2.5. comment 5). The scale was based on the studies involving Krumhansl (see Chapter 3) and, interestingly, in none of these studies is there an explanation for using a scale from 1 to 7. Overall, this method of judging the samples was found to be somewhat confusing.

Test 1 was considered frustrating by one subject due to the way the play and pause buttons functioned, so his frustration ranking was not based on the method only but also on the user interface. This subject would have liked the user interface to have a replay button to make it easier to listen to the samples again. These remarks are very important and had to be taken into consideration in the design of the user interface for the large scale experiment.

Test 2 was considered too long, and one subject commented that he thought his concentration started to slip, "leading to repeated listens of the samples" (A2.6. comment 4). One subject noted that 66 trials is a lot to go through, and having to listen to so many A-based chords is monotonous. This subject also claimed that she had started to doubt her perfect pitch during the test. Moreover, there were comments on ambiguity regarding the task of judging how well the tones in the samples fit together musically. According to the comments (cf. appendix A2.6. comments) some of the subjects felt unsure about their judgements in this test. It is possible that the length and number of trials caused fatigue and uncertainty in the subjects.

Test 3 gained some positive comments except for its user interface. One subject commented that he liked the idea of this test the most, but the user interface made it the most difficult test. This subject also recommended a user interface with movable "drag'n'drop" (cf. A2.7. comment 3) boxes to make the test easier. Another subject expressed similar thoughts on the user interface (ibid: comment 4). Having only twelve samples to concentrate on was considered a positive thing, and the task of ordering them was found to be more interesting than rating stimuli one by one or by comparing pairs (cf. ibid: comment 1).

The average ratings for easiness, boredom, frustration and time taken to complete the test along with percentages of 'yes'-answers to the most important questions of the test questionnaires are found in Table 2 below. More detailed results of the questionnaires are found in appendix A2. All of the eight subjects had completed all of the tests, so there is no data in Table 2 for how many subjects had quit the test.

**Table 2**

| **Test** | *Felt like quitting* | *Easiness* | *Boredom* | *Frustration* | *Would participate** | *Would complete** | *Time taken* |
|---|---|---|---|---|---|---|---|
| *Test 1* | 0.0% | 3.63 | 2.63 | 2.50 | 62.5% | 37.5% | 5min 25sec |
| *Test 2* | 37.5% | 4.13 | 3.25 | 3.38 | 50.0% | 12.5% | 13min 27sec |
| *Test 3* | 0.0% | 4.38 | 2.38 | 2.75 | 75.0% | 62.5% | 7min 21sec |

**Would participate if found the test on an Internet forum*
**Would complete the test if found it on an Internet forum*

According to these ratings the method used in test 1 would be a fairly suitable one. Test 1 did not take very long, and it was not considered the most boring or frustrating. On the other hand the task was not rated to be as easy as the tasks in the other tests. Along with the comments questioning the meaningfulness of the rating scale of 1 to 7 in test 1, this method cannot be considered the best of these three for the purpose of investigating the perception of consonance.

Only test 2 had caused subjects to answer that they had felt like quitting the test. The task of comparing pairs was found to be easy, but the test was considered the most boring and frustrating of the three tests. It also took the longest. These findings imply that it is problematic to use paired comparisons for online testing because boredom and frustration caused by a long duration are likely to cause a lot of dropout.

Test 3 was rated to have the easiest task and to be the least boring. In the general questionnaire on the entire pilot experiment, test 2 and 3 were tied for the easiest test (cf. appendix A2.8.). Test 3 was rated more frustrating than test 1, but this might be due to the user interface of test 3, which was criticised in the comments. Test 3 also did not take very long compared to test 2. In addition, most of the subjects thought that they would participate in test 3 and complete it if they found it on an Internet forum. In the general questionnaire test 3 was also the most liked test, but three subjects still ranked it as the test they would be most likely to quit if they found it on an Internet forum. Also in the general comments on the pilot experiment, test 3 was commented to be the

preferred test, but its user interface was again criticised (see A2.8. comment 7). Based on the results of the pilot experiment, the most suitable method from these three tested methods is rank order paradigm. However, a better user interface is needed for the test.

The results of the pilot experiment implicate that some improvements and changes must be made to the methodology before it can be used for a large scale experiment. Regarding the questionnaire, the questions on musical background could be more specific as was already mentioned. Whether a subject has training in music theory or ear training should be asked separately, and the type of music training should also be asked. Based on the many comments on the ambiguity of *fitting together musically*, the presentation of the task has to be changed. It would be best to simply use the word 'musical consonance' in the task presentation. This way, musical subjects would instantly understand the purpose of the task. By including a simple definition for musical consonance, the task would be understandable to non-musical subjects also. Test 3 of the pilot experiment had no instructions on the test page about how to rank the stimuli, and, as a result, one of the subjects used the ranking in opposite order. Therefore, instructions should be visible on the test page too to ensure that the subjects remember how they are supposed to rank the stimuli.

# 6. THE DESIGN OF THE LARGE SCALE EXPERIMENT

This chapter describes the design of a large scale online listening test experiment for investigating the perception of harmony. The formal research question of the large scale experiment will be described along with the hypotheses. Using the results of the pilot experiment, the test method was chosen and a user interface was designed.

## 6.1. Research question and hypotheses for the experiment

The first step in designing a listening test experiment is defining the research question that is to be answered by the experiment's results (Bech & Zacharov 2006: 17). This study is concerned with the design of an online listening test experiment for investigating the musical consonance of intervals in relation to power chords. A power chord is a harmonic structure used often in heavy metal music, typically consisting of a fifth and sometimes an octave played with distortion (cf. Lilja 2009: 102). The reason for using a power chord and a tone to form the intervals instead of simply using intervals between two tones (*dyads*) is that the power chord has such an important role in heavy metal (ibid). Investigating the perception of harmony with power chords can thus produce results which are more relevant for heavy metal harmony than results obtained with dyads. The goal is to find out which intervals are considered most musically consonant in relation to the power chord. The reasons for developing an online listening test are the possibility of a larger and more diverse sample and the fact that the internal validity of a music listening test is not significantly compromised by being conducted online. The dependent variable is the musical consonance of the intervals. Essentially, the purpose of the experiment is to gather data similar to the data gathered by Krumhansl and Kessler (1982: 343), and Krumhansl and Shepard (1979: 586), except that the data will be gathered by using an ordinal scale ranking system for the listening tests.

Intervals are traditionally divided into consonant and dissonant intervals. Consonant intervals include minor and major thirds, minor and major sixths, and all perfect

intervals. Dissonant intervals include minor and major seconds, minor and major sevenths, and all augmented and diminished intervals. The perfect fourth is an exception to this division, as it can be considered either a dissonance or a consonance depending on its context (Salmenhaara 2005: 25; Piston 1962: 6). The results of Krumhansl and Kessler (1982: 343) are in accordance with the traditional division of intervals into consonances and dissonances as the intervals traditionally considered consonant were judged to fit better into the context than the intervals traditionally considered dissonant. The results of Krumhansl and Shepard (1979: 586) and McDermott et al. (2010: 1036) are similar: the intervals traditionally considered consonant were ranked more consonant than the intervals traditionally considered dissonant. Despite this, in none of the aforementioned studies are different intervals divided into consonances and dissonances. Lilja (2009: 134) argues that the division of intervals into consonances and dissonances is "at best, arbitrary". Therefore, also in the context of this study, it is more useful to view consonance as a continuum from most consonant to least consonant (i.e. most dissonant) instead of arbitrarily dividing the intervals into consonances and dissonances. The goal is simply to establish an order from the most consonant interval to the least consonant.

Lilja (2009: 112–113) argues that distortion has significant effects on the harmonic content of power chords and that power chords contain harmonics that a pure fifth interval does not contain when played without distortion. Therefore, it is necessary to investigate whether the intervals conform to the traditional concepts of consonance and dissonance (cf. Salmenhaara 2005: 25) and to the results of studies such as the one by Krumhansl and Kessler (1982: 343). If the consonance of intervals with distortion differs from the consonance of intervals without distortion, it is evident that many traditional concepts of harmony cannot be applied to heavy metal music. However, if the perception of power chords is found to resemble the perception of some musical context used in earlier experiments, this information can be used to interpret power chords as similar to that context. For example, if the results of the experiment resemble the results of the major context test by Krumhansl and Kessler (1982: 343), the power chord can be considered to create a major context harmonically.

In listening test experiments it is necessary to define a hypothesis. There are two possible hypotheses for the experiment with intervals and power chords. The results can be similar to the ratings of intervals in a major context obtained in previous studies (cf. Krumhansl & Kessler 1982; Krumhansl & Shepard 1979) because a power chord is likely to be perceived as a major context due to the major third in its overtone series (Lilja 2009: 113–114). This hypothesis is supported by the results of the pilot experiment although those results are not conclusive due to the small sample and lack of randomisation within the tests. The other possible hypothesis is that distortion has such a significant effect on the perceived consonance of intervals that the results differ greatly from the traditional classification of consonance and dissonance (cf. Salmenhaara 2005: 25; Piston 1962: 6) and from the results of experiments without distortion (cf. Krumhansl & Kessler 1982: 343; McDermott et al. 2010: 1036).

Returning to Ekman and Sjöberg's (1965: 452) question ("What is being measured?") and to the comments on the task in the pilot experiment, it is evident that the subjects must be presented with a different task. The goal is to measure the perceived musical consonance of intervals. Therefore, it is best to use the word *consonance* in describing the task. The ranking method to be used in the experiment will be rank-order paradigm, in which the subjects put the samples in order from most consonant to least consonant. Musical subjects can be assumed to know what musical consonance is, whereas non-musical subjects can be given a brief explanation about consonance. For example, the task could be phrased in the following way: "Your task is to put the samples in order from most consonant to least consonant. In this experiment consonance refers to the musical quality of multiple tones forming a pleasant combination. In other words, consonance means how well the different tones sound together musically."

## 6.2. The user interface

Test 3 of the pilot experiment was generally liked by the subjects, but its user interface was criticised. The user interface was considered both difficult to use and frustrating. A better user interface was designed with two very important aspects of user interface

design in mind*:* usability and clarity. These are both important factors in reducing dropout and bias.

According to Rogers et al. (2012: 19), "*[u]sability* refers to ensuring that interactive products are easy to learn, effective to use, and enjoyable from the user's perspective". In the comments of test 3 of the pilot experiment, two subjects recommended a user interface with draggable elements for ordering the samples. This type of interface is very easy to learn because nowadays most operating systems have such interfaces with draggable windows representing the interfaces of different programs. It is also effective as such an interface can make it easier for the subjects to remember which samples they have compared to each other, by dragging them next to each other, for example. Having a familiar type of user interface probably also makes the interface more enjoyable from the user's perspective. One of the goals of creating a user interface that is usable is to make one whose usage does not cause frustration, which is something that can be caused by an interface that makes a simple task complicated (ibid: 135). This is exactly what happened with test 3 of the pilot experiment: the simple task of ordering samples became frustrating due to an interface with very low usability.

The user interface for the listening test in the large scale experiment designed here will consist of boxes that can be dragged using the mouse. Each of these boxes represents one of the samples and contains a button that can be used to play the sample. The boxes will be placed in random locations inside a limited area. At the bottom of the screen, there is a grid on which the boxes can be ordered from most consonant to least consonant. Once the subject has ordered all the samples on the grid the subject can press the 'submit' button. The grid will have the anchors *most consonant* and *least consonant* to remind the subjects about how to use the scale. This should deter the subjects from using the scale in opposite order. Also a button that shows the instructions again will be included so that the subjects can review them if they wish. This is because of the possibility that a subject does not read the instructions properly before starting the test, so they should be available during the test also.

The boxes representing the samples contain a triangle that is typically associated with the *play* function in music players and audio playback programs. The play symbol turns into a filled in black box so that it is clearly visible which sample is playing. The box symbol was chosen because it is typically associated with the *stop* function in music players. The meaning of these symbols will be described in the instructions. Pressing 'stop' on a sample returns the sample to its beginning because it was noted in the comments of the pilot experiment that replaying the samples from the beginning would be preferable to having them continue from the point in which they were paused. Pressing 'play' in one box also stops any other sample that is playing. The purpose of having functions such as these is to avoid situations where the subject does not know which sample is playing and to avoid the possibility of having multiple samples playing simultaneously. The aim is to make the user interface easier to use and to reduce frustration. The boxes have a light grey background and borders. Using both borders and a contrasting background colour have been shown to aid in finding text on a screen (Weller 2004), and it can also help to make this interface clearer and easier to use. A sketch of this interface can be seen below in Figure 2, which demonstrates a situation where one sample is placed on the grid and another sample is playing.

The process of participating in the experiment is depicted using a *storyboard* (cf. Rogers et al. 2012: 418) in Figure 3. The first page of the experiment contains an introductory text explaining the purpose of the experiment in sufficient detail and the questionnaire on musical background. The questionnaire is placed in the beginning to reduce dropout (cf. Reips 2002a: 243). The second page of the experiment contains the instructions for the listening test and has a volume calibration sample so that the subjects can set their volume to a comfortable level before the actual test. The actual listening test is found on the third page. Once the subject presses the 'submit' button, the browser moves to the fourth page, where the subject is thanked for participating and told that the results have been submitted and that the subject can close his browser window.

Figure 2

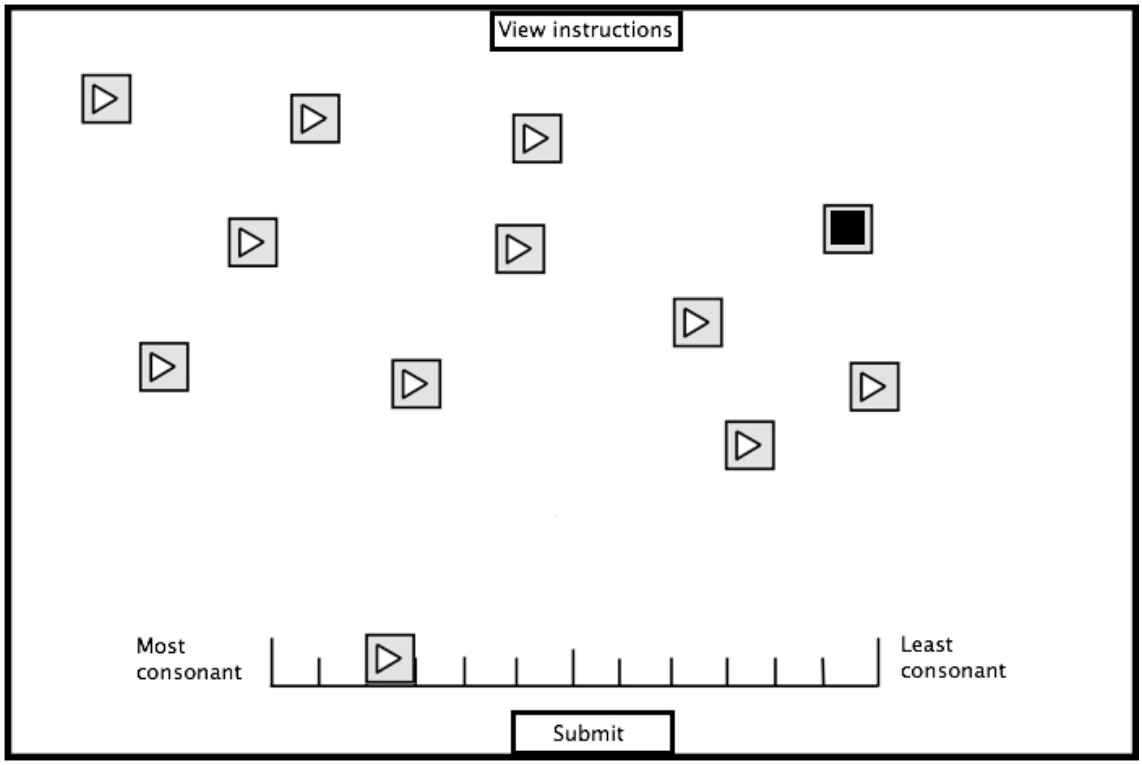
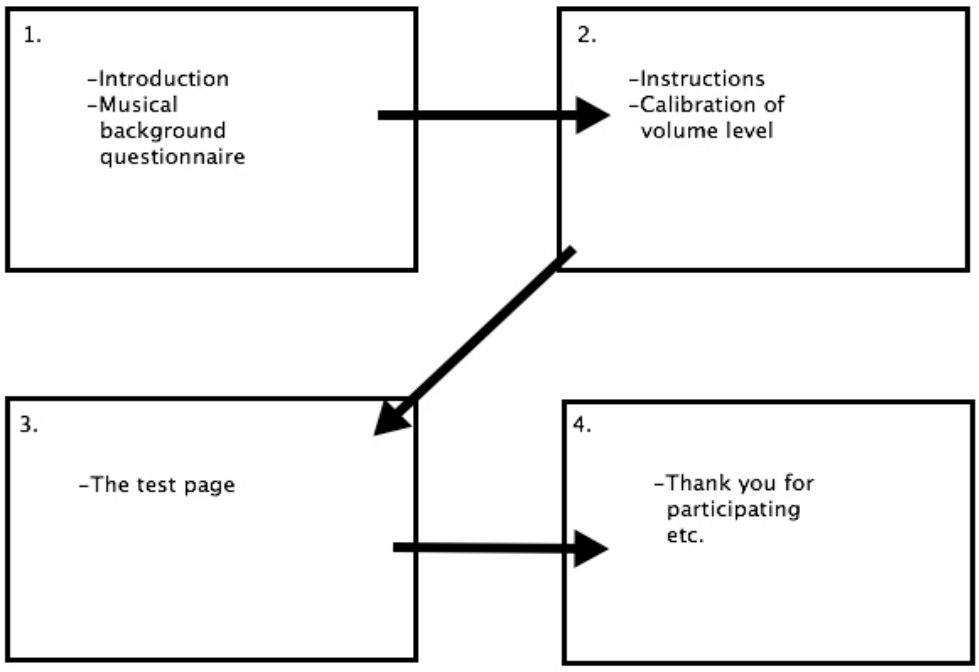



Figure 3



To avoid bias caused by the placement of the sample boxes, the positions of the sample boxes are randomised separately for each subject. The boxes will be placed above the grid, leaving some empty space above it, so that the subjects have room to move the boxes around before placing them onto the grid. There will be a post-selection procedure to ensure that unusable results are not saved. The questionnaire on musical background will be checked to make sure that the answers make sense. For example, if a subject claims to have more years of musical training than his age is, the results will be discarded. In the test it will be necessary to keep track of which samples the subject has listened to. If the subject has not listened to all of the samples, the results will be discarded. There will also be a timer to make sure that the subject has spent enough time on the test to indicate that he has made his judgements properly. The same samples will probably be used during the large scale experiment because they were not criticised by the subjects in the pilot experiment.

## 6.3. Recruiting subjects

The subjects will be recruited for the experiment using e-mail lists and Internet forums among other media. There will be no reward for the subjects, so they will participate purely out of interest towards the experiment. While this can lead to a smaller sample, it is also likely to reduce the possibility of multiple entries and other dishonest answers. A volunteer sample can be biased because those people who are interested in the experiment might not necessarily represent any population very well. A large and diverse sample of Internet users is still very likely to be more representative than the samples of university students traditionally used for music perception experiments (cf. chapter 3). Aiming for a large sample hopefully brings diversity into the sample, which can then be divided into different categories using the results from the musical background questionnaire. In a music perception experiment, the musical background of the subjects is the most important aspect, and as long as there are enough musical and non-musical subjects participating, the lack of diversity in socioeconomic status etc. is not significant.

The volunteer subjects are very likely to be amateur or professional musicians, and using music-related Internet forums is potentially one of the best places to recruit musical subjects. Recruiting non-musical subjects can be more difficult. E-mail lists for different university student associations can be used to obtain non-musical subjects. It is also important to conduct a parallel experiment in a laboratory setting using the same Internet experiment application. This way the potential effects of uncontrolled test environments can be analysed. Overall, the main recruitment strategy is to obtain as large and diverse a sample as possible. Having a large and diverse sample is, in fact, one of the greatest strengths of online listening tests and also the best way to reduce the effects of uncontrolled test environments.

# 7. CONCLUSIONS

Overall, Internet experimenting is a fairly new method of conducting experiments. The goal of this study was to create a good and valid design for an online listening test for investigating the perception of heavy metal harmony. Methodology was gathered from sources describing the general methodology of listening tests, from laboratory based music perception experiments and from online listening tests experiments. A pilot experiment was conducted to obtain more detailed data on the suitability of different methods for a large scale online experiment.

Considering the studies on pitch and music perception that were presented in the third chapter of this study, it is evident that a lot can be gained from conducting music perception experiments online. Music perception experiments in laboratory settings have often had very small samples that have lacked in diversity as well, whereas with online experiments it is easier to obtain large and diverse samples. Online experiments have potential issues including the lack of control, i.e. internal validity, and possibly skewed samples. It has been shown by Disley et al. (2006) that online experiments can produce reliable results even in tests on the perception of timbre, which typically require strict control over the test environment and equipment. Computers are used by many for music listening, and subjects can participate in the experiment at home. This can result in a very natural and familiar test environment from the subject's perspective. The samples that can be obtained online are also likely to be far more diverse than the samples that have been used in laboratory based listening tests. Therefore, it is reasonable to claim that the possibilities of large and diverse samples easily make up for the loss of some control in online music perception experiments.

Many of the basic aspects of listening tests are just as relevant for online tests as laboratory based tests. The order of stimuli must be randomised or balanced to avoid time order errors. Visual and expectation bias can be caused by the user interface of an online test, for which reason the design of a user interface is an integral part of designing the whole test procedure. Bad user interfaces can lead to frustration that potentially contributes to greater dropout rates and harms the judgement making of the

subjects.

The results of the pilot experiment implicated that it is important to make the test easy to participate in and easy to complete. Also the task of making judgements should be meaningful to the subjects. The test procedure must be interesting and short enough so that the subjects actually want to complete the test. Submitting results should not require more than pressing a ‘submit’ button on a website. Using more complicated systems for submitting the results, such as html's mailto-method, will probably cause potential subjects to not even participate. Rank-order paradigm was selected as the test method for the large scale experiment, and an improved user interface was designed based on the suggestions made by the subjects. Once the first working version of the large scale experiment is ready, further pilot experimenting will be necessary before launching the actual experiment.

The detailed technical aspects of implementing an online listening test application were omitted from this study as they are beyond the scope of this study. Designing an online listening test requires pilot experimenting and careful design as there is not yet any established methodology for conducting music perception experiments online. The results of the pilot experiment in this study, and this study overall, serve as a starting point for developing an online listening test for investigating the perception of heavy metal harmony. Hopefully, this study can also serve as a stepping stone for musicologists planning to use online listening tests in their research.

References

Andrews, Dorine & Nonnecke, Blair & Preece, Jennifer 2003. "Electronic Survey Methodology: A Case Study in Reaching Hard-to-Involve Internet Users." *International Journal of Human-Computer Interaction* 16(2): 185–210.

Attneave, Fred & Olson, Richard K. 1971. "Pitch as a medium: a new approach to psychophysical scaling." *The American Journal of Psychology* 84(2): 147–166.

Balch, Charles V. 2010. *Internet Survey Methodology.* Newcastle, UK: Cambridge Scholars Publishing.

Bech, Søren & Zacharov, Nick 2006. *Perceptual Audio Evaluation – Theory, Method and application.* West Sussex, England: John Wiley & Sons Ltd.

Bharucha, Jamshed & Krumhansl, Carol L. 1983. "The representation of harmonic structure in music: Hierarchies of stability as a function of context." *Cognition* 13: 63–102.

Bigand, E. & Poulin-Charronnat, B. 2006. "Are we 'experienced listeners'? A review of the musical capacities that do not depend on formal musical training." *Cognition* 100(1): 100–130.

Bowers, Diane K. 1998. "FAQs on Online Research." *Marketing Research* (Winter 1998/Spring 1999): 45–48.

Brattico, E. & Pallesen K. J. & Varyagina, O. & Bailey, C. & Anourova, I. & Järvenpää, M. & Eerola, T. & Tervaniemi, M. 2008. "Neural Discrimination of Nonprototypical Chords in Music Experts and Laymen: An MEG Study." *Journal of Cognitive Neuroscience* 21(11): 2230–2244.

Christian, Leah Melani & Dillman, Don A. 2004. "The influence of graphical and symbolic language manipulations on responses to self-administered questions." *Public Opinion Quarterly* 68(1): 57–80.

Cox, Trevor J. 2007. "Bad vibes: an investigation into the worst sounds in the world." Presented at the *19th International Congress on Acoustics.*

Cuddy, Lola L. 1971. "Absolute judgement of musically-related pure tones." *Canadian Journal of Psychology* 25(1): 42–55.

Cuddy, Lola L. & Cohen, Annabel J. 1976. "Recognition of Transposed Melodic Sequences." *Quarterly Journal of Experimental Psychology* 28(2): 255–270.

Cuddy, Lola L. & Lyons, H. I. 1981. "Musical pattern recognition: a comparison of listening to and studying tonal structures and tonal ambiguities." *Psychomusicology* 1(2): 15–33.

David, H. A. 1988 [1963]. *The Method of Paired Comparisons.* London: Charles Griffin & Company Limited.

Deutsch, Diana 1972. "Mapping of Interactions in the Pitch Memory Store." *Science* 175(3): 1020–1022.

Disley, Alastair C. & Howard, David M. & Hunt, Andy D. 2006. "Timbral description of musical instruments." In *Proceedings of the 9th International Conference on Music Perception and Cognition*: 61–68.

Dowling, W. J. & Fujitani, Diane S. 1971. "Contour, Interval, and Pitch Recognition in Memory for Melodies." *The Journal of the Acoustical Society of America* 49(2): 524–531.

Dowling, W. J. 1978. "Scale and Contour: Two Components of a Theory of Memory for Melodies." *Psychological Review* 85(4): 341–354.

Duerst, M. & Masinter, L. & Zawinski, J. 2010. "The 'mailto' URI Scheme" *RFC 6068.* Internet Engineering Task Force. https://tools.ietf.org/html/rfc6068 (retrieved March 18th, 2015)

Egermann, Hauke & Nagel, Frederik & Kopiez, Reinhard & Altenmüller, Eckart 2006. "Online measurement of emotional musical experiences using internet-based methods – An exploratory approach." In *Proceedings of the 9th International Conference on Music Perception and Cognition*: 178–183.

Ekman, Gösta & Sjöberg, Lennart 1965. "Scaling." *Annual Review of Psychology* 16: 451–474.

Fortson, Beverly L. & Scotti, Joseph R. & Del Ben, Kevin S. & Chen, Yi-Chuen 2006. "Reliability and Validity of an Internet Traumatic Stress Survey with a College Student Sample." *Journal of Traumatic Stress* 19(5): 709–720.

French-Lazovik, Grace & Gibson, Curtis L. 1984. "Effects of Verbally Labeled Anchor Points on the Distributional Parameters of Rating Measures." *Applied Psychological Measurement* 8(1): 49–57.

Goldsmiths 2015. "The Goldsmiths Musical Sophistication Index." http://www.gold.ac.uk/music-mind-brain/gold-msi/ (retrieved February 6th,

2015)

Hewson, Claire & Yule, Peter & Laurent, Dianna & Vogel, Carl 2003. *Internet research methods: A Practical guide for the Social and Behavioural Sciences.* London, UK: SAGE Publications.

Honing, Henkjan 2006. "Evidence for Tempo-Specific Timing in Music Using a Web-Based Experimental Setup." *Journal of Experimental Psychology: Human Perception and Performance* 32(3): 780–786.

Honing, Henkjan & Ladinig, Olivia 2006a. "The effect of exposure and expertise on timing judgements in music: Preliminary results." In *Proceedings of the 9th International Conference on Music Perception and Cognition*: 80–85.

Honing, Henkjan & Ladinig, Olivia 2006b. Questionnaire in an online listening test. http://www.mcg.uva.nl/drafts/EEE-online/EEE-questionnaire.html (retrieved February 24th, 2015)

Honing, Henkjan 2007. "Is expressive timing relational invariant under tempo transformation?" *Psychology of Music* 35(2): 276–285.

Honing, Henkjan & Ladinig, Olivia 2008. "The Potential of the Internet for Music Perception Research: A Comment on Lab-Based Versus Web-Based Studies." *Empirical Musicology Review* 3(1): 4–7.

Honing, Henkjan & Ladinig, Olivia 2009. "Exposure Influences Expressive Timing Judgements in Music." *Journal of Experimental Psychology: Human Perception and Performance* 35(1): 281–288.

Hynninen, Jussi 2001. *A software-based system for listening tests.* Master's thesis. Helsinki University of Technology.

Im, Eun-Ok & Chee, Wonshik 2011. "Quota Sampling in Internet Research. Practical Issues." *CIN: Computers, Informatics, Nursing* 29(7): 381–385.

Internet live stats 2015. "Internet users." http://www.internetlivestats.com/watch/internet-users/ (retrieved February 2nd, 2015).

Internet World Stats 2015. "Internet usage statistic." http://www.internetworldstats.com/stats.htm (retrieved February 2nd, 2015).

ITU-R 2014. *Recommendation BS.1116-2. Methods for the subjective assessment of small impairments in audio systems.* International Telecommunications Union.

ITU-T 1996. *Recommendation P.800. Methods for the subjective determination of transmission quality.* International Telecommunications Union.

Janata, Petr & Birk, Jeffrey L. & Tillmann, Barbara & Bharucha, Jamshed J. 2003. "Online Detection of Tonal Pop-Out in Modulating Contexts." *Music Perception* 20(3): 283–305.

Kameoka, Akio & Kuriuagawa, Mamoru 1969a. "Consonance Theory Part I: Consonance of Dyads." *The Journal of the Acoustical Society of America* 45(6): 1451–1459.

Kameoka, Akio & Kuriuagawa, Mamoru 1969b. "Consonance Theory Part II: Consonance of Complex Tones and Its Calculation Method." *The Journal of the Acoustical Society of America* 45(6): 1460–1469.

Kendall, M. G. & Babington Smith B. 1940. "On the Method of Paired Comparisons." *Biometrika* 31(3): 324–345.

Kosta, Katerina & Song, Yading & Fazekas, György & Sandler, Mark B. 2013. "A study of cultural dependence of perceived mood in Greek music." Presented at the *14th International Symposium on Music Information Retrieval*: 317–322.

Krumhansl, Carol L. & Shepard, Roger N. 1979. "Quantification of the Hierarchy of Tonal Functions Within a Diatonic Context." *Journal of Experimental Psychology* 5(4): 579–594.

Krumhansl, Carol L. 1979. "The Psychological Representation of Musical Pitch in a Tonal Context." *Cognitive Psychology* 11: 346–374.

Krumhansl, Carol L. & Kessler, Edward J. 1982. "Tracing the Dynamic Changes in Perceived Tonal Organization in a Spatial Representation of Musical Keys." *Psychological Review* 89(4): 334–368.

Krumhansl, Carol L. & Keil, Frank C. 1982. "Acquisition of the hierarchy of tonal functions in music." *Memory & Cognition* 10(3): 243–251.

Krumhansl, Carol L. & Bharucha, Jamshed J. & Kessler, Edward J. 1982. "Perceived Harmonic Structure of Chords in Three Related Musical Keys." *Journal of Experimental Psychology: Human Perception and Performance* 8(1): 24–36.

Krumhansl, Carol L 1990. *Cognitive Foundations of Musical Pitch.* Oxford: Oxford University Press.

Kruskal, J. B. 1964. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika* 29(1): 1–27.

Kurose, James F. & Ross, Keith W. 2013. *Computer networking. A Top-Down Approach.* 6th edition. Essex, England: Pearson.

Lawless, Harry T. & Heymann, Hildegarde 1998. *Sensory Evaluation of Food: Principles and Practices.* New York: Chapman & Hall.

Levelt, W. J. M. & van de Geer, J. P. & Plomp R. 1966. "Triadic comparisons of musical intervals." *The British Journal of Mathematical and Statistical Psychology* 19(2): 163–179.

Lilja, Esa 2009. *Theory and Analysis of Classic Heavy Metal Harmony.* Vantaa: IAML Finland.

Marques, Carlos & Moreno, Sylvain & Castro, São Luís & Besson, Mireille 2007. "Musicians Detect Pitch Violation in a Foreign Language Better Than Nonmusicians: Behavioral and Electrophysical Evidence." *Journal of Cognitive Neuroscience* 19(9): 1453–1463.

McDermott, Josh H. & Lehr, Andriana J. & Oxenham, Andrew J. 2010. "Individual Differences Reveal the Basis of Consonance." *Current Biology* 20(11): 1035–1041.

McGraw, Kenneth O. & Tew, Mark D. & Williams, John E. 2000. "The integrity of web-delivered experiments: Can You Trust the Data?" *Psychological Science* 11(6): 502–506.

Novitski, Nikolai & Tervaniemi, Mari & Huotilainen, Minna & Näätänen, Risto 2004. "Frequency discrimination at different frequency levels as indexed by electrophysiological and behavioral measures." *Cognitive Brain Research* 20: 26–36.

Novitski, Nikolai 2006. *Pitch discrimination in optimal and suboptimal acoustic environments: electroencephalographic, magnetoencephalographic, and behavioral evidence.* Doctoral dissertation. University of Helsinki.

Nummenmaa, Lauri 2011. *Käyttäytymistieteiden tilastolliset menetelmät.* Helsinki: Sanoma Pro.

Piston, Walter 1962 [1941]. *Harmony.* New York: W.W. Norton & Company.

Plomp, R. & Levelt, W. J. M. 1965. "Tonal Consonance and Critical Bandwidth." *Journal of the Acoustical Society of America* 38: 548–560.

Reips, Ulf-Dietrich 2002a. "Internet-Based Psychological Experimenting. Five Dos and Five Don'ts." *Social Science Computer Review* 20(3): 241–249.

Reips, Ulf-Dietrich 2002b. "Standards for Internet-Based Experimenting." *Experimental Psychology* 49(4): 243–256.

Rogers, Yvonne & Sharp, Helen & Preece, Jenny 2012 [2011]. *Interaction design: beyond human-computer interaction.* West Sussex, UK: John Wiley & Sons Ltd.

Salmenhaara, Erkki 2005 [1968]. *Sointuanalyysi.* Helsinki: Otava.

Schellenberg, E. Glenn & Trehub, Sandre E. 1994a. "Frequency ratios and the perception of tone patterns." *Psychonomic Bulletin & Review* 1(2): 191–201.

Schellenberg, E. Glenn & Trehub, Sandre E. 1994b. "Frequency ratios and the discrimination of pure tone sequences." *Perception & Psychophysics* 56(4): 472–478.

Schellenberg, E. Glenn & Trainor, Laurel J. 1996. "Sensory consonance and the perceptual similarity of complex-tone harmonic intervals: Tests of adult and infant listeners." *Journal of the Acoustical Society of America* 100(5): 3321–3328.

Scriven, Frances 2005. "Two types of sensory panels or are there more?" *Journal of Sensory Studies* 20: 526–538.

Simsek, Zeki & Veiga, John F. 2001. "A Primer on Internet Organizational Surveys." *Organizational Research Methods* 4(3): 218–235.

Smith, Michael A. & Leigh, Brant 1997. "Virtual subjects: Using the Internet as an alternative source of subjects and research environment." *Behavior Research Methods, Instruments & Computers* 29(4): 496–505.

Song, Yading & Dixon, Simon & Pearce, Marcus & Halpern, Andrea R. 2013. "Do Online Social Tags Predict Perceived of Induced Emotional Responses to Music?" Presented at the *14th International Symposium on Music Information Retrieval*: 89–94.

Stevens, S. S. & Volkmann J. & Newman, E. B. 1937. "A Scale for the Measurement of the Psychological Magnitude of Pitch." *Journal of the Acoustical Society of America* 8(3): 185–190.

Stone, Herbert & Bleibaum Rebecca N &; Thomas, Heather A. 2012. *Sensory Evaluation Practices.* Academic Press.

Terhardt, Ernst 1984. "The Concept of Musical Consonance: A Link between Music and Psychoacoustics." *Music Perception* 1(3): 276–295.

Thurstone, L. L. 1927a. "A law of comparative judgement." *Psychological Review* 34: 273–286.

Thurstone, L. L. 1927b. "Psychophysical analysis." *American Journal of Psychology* 38: 368–369.

Toole, Floyd E. & Olive, Sean E. 1994. "Hearing is Believing vs. Believing is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things." Presented at the *97th AES Convention*, Paper 3894.

Trehub, Sandra E. 2001. "Musical Predisposition in Infancy." *Annals of the New York Academy of Sciences* 930: 1–16.

Virtala, Paula 2015. *The neural basis of Western music chord categorisations – effects of development and music expertise.* Doctoral dissertation. University of Helsinki.

W3Techs 2015. "Usage of JavaScript for websites." http://w3techs.com/technologies/details/cp-javascript/all/all (retrieved February 2nd, 2015)

Wallentin, M. & Nielsen, A. H. & Friis-Olivarius, M. & Vuust, C. & Vuust, P. 2010. "The Musical Ear Test, a new reliable test for measuring musical competence." *Learning and Individual Differences,* 20(3): 188–196.

Weller, Donnelle 2004. "The effects of contrast and density on visual web search." *Usability News* 6(2). http://psychology.wichita.edu/surl/usabilitynews/62/density.htm (retrieved March 16th, 2015)

Wright, Matthew 2008. *The shape of an instant: Measuring and modeling perceptual attack time with probability density functions.* Doctoral dissertation. Stanford University.

Zielinski, Slawomir 2006. "On Some Biases Encountered in Modern Listening Tests." Presented at the *Spatial Audio & Sensory Evaluation Techniques Workshop.* Guildford, UK 6–7 April.

# Appendix A1: Pilot experiment materials

## A1.1. General instructions

**Welcome to the pilot experiment!**

I am a student of musicology at the University of Helsinki and I am working on my master's thesis. The subject of my thesis is the design of an online listening test for investigating the perception of harmony in music. This is a pilot experiment, consisting of five separate parts. The results of this experiment will be used in the design of a large scale online listening test. The results you submit will be handled anonymously, so your name and e-mail address will not be included in my thesis. By participating in this experiment, you give me permission to use your answers in my thesis. The purpose of this test is not to test you, but to test different experimental arrangements, so even if you have negative feedback about the experiment, it is extremely valuable information.

General instructions:
This experiment consists of five parts: three listening tests and two questionnaires. There is also a short questionnaire at the end of each listening test. I hope that you will try to complete all of the listening tests, but if you feel like you don't want to complete one of the tests, you can just move on to the questionnaire at the end of that test and submit incomplete results. I also hope that you will be able to submit all your results by the **5th of March**. The experiment is implemented as a set of html-files, which should open in your default Internet browser. For sending the results it is important that you have an e-mail program configured on your computer (such as Microsoft Outlook, Apple Mail, or Mozilla Thunderbird). When you submit results, your e-mail program will open a new message with the results and my e-mail address, so that you only have to click "Send". No message title is required. If you prefer to send the results in some other way, you can simply copy–paste them from the message field and send them to otso.bjorklund@helsinki.fi. You can also contact me in the case of technical problems using the same e-mail address.

**The different parts of the experiment should be completed in the following order:**

1. Musical background
2. Test 1
3. Test 2
4. Test 3
5. Questionnaire

All the parts of the experiment are contained in the folder "The Experiment".
It is best that you take a break between the parts, and you can even complete different parts on different days. Each of the parts contains more specific instructions concerning that part.

*Otso Björklund*

A1.2. Questionnaire on musical background

## Musical background.

Please answer the following questions about yourself and your musical background.

### 1. General information

E-mail address: (this will be removed from the answers before publication)

Age:

### 2. Listening to music

How many hours do you spend listening to music weekly? (estimate)

Please try to distribute this time in percent:

% Classical
% Jazz
% Rock
% Pop
% Heavy metal
% other

What kind of a music listener do you consider yourself?

- 1 (Casual listener)
- 2
- 3
- 4
- 5 (Enthusiastic listener)

### 3. Musical background

Have you ever received training in a musical instrument (or voice)?

- Yes
- No

Which instruments have you played (if any)?

Have you ever received training in music theory / ear training?
- ○ Yes
- ○ No

At what age did you begin taking music lessons? (leave empty if you have never taken music lessons)

Total duration of musical training in years?

How many hours per week do you spend playing an instrument or practicing music (e.g. theory, ear training)?

Can you recognize musical intervals by ear?
- ○ 1 (Not at all)
- ○ 2
- ○ 3
- ○ 4
- ○ 5 (With very good accuracy)

Do you have perfect pitch?
- ○ Yes
- ○ No

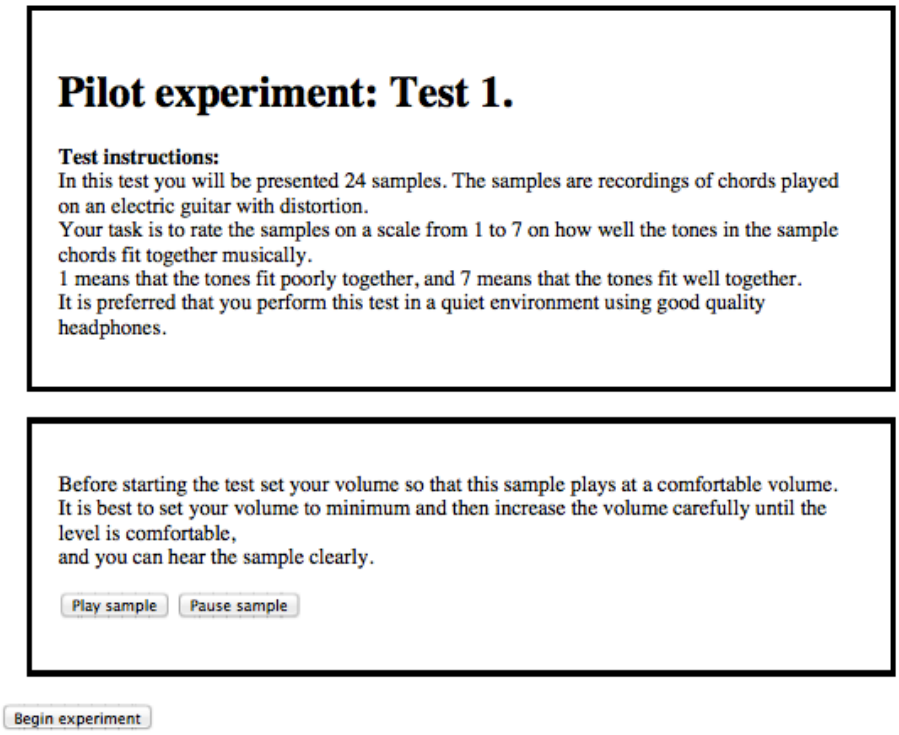
Are you any of the following? (You can check multiple boxes)
- ☐ Professional musician
- ☐ Musicologist / Music theorist
- ☐ Music teacher
- ☐ Instrument teacher
- ☐ Music producer / engineer
- ☐ None of these apply to me

Additional comments (if you have any):

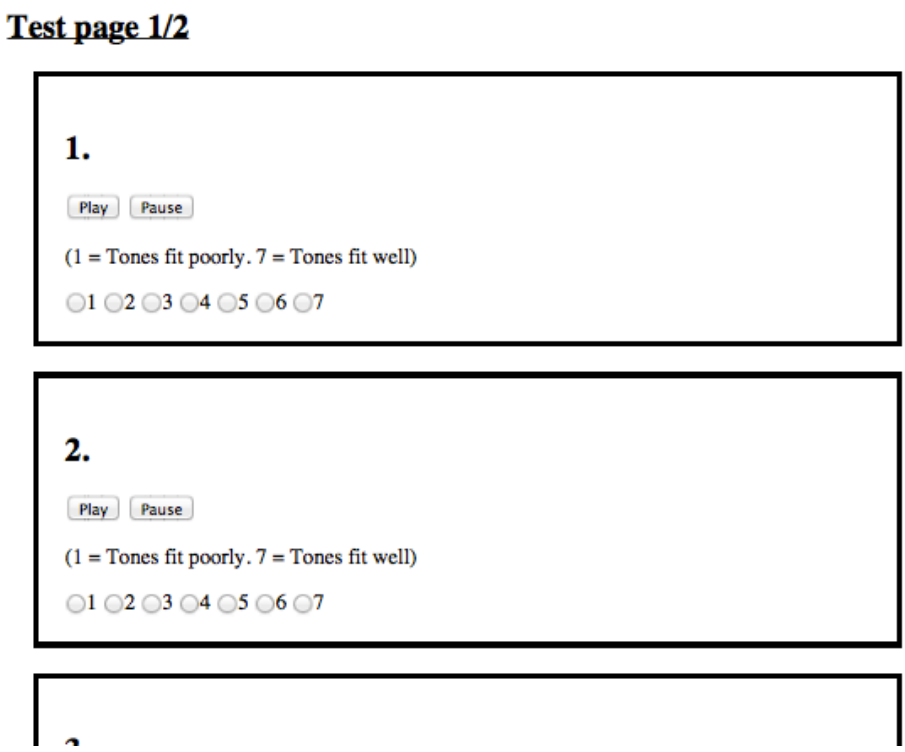
Once you have clicked Submit and sent the e-mail containing the results you can close this browser window.

Submit

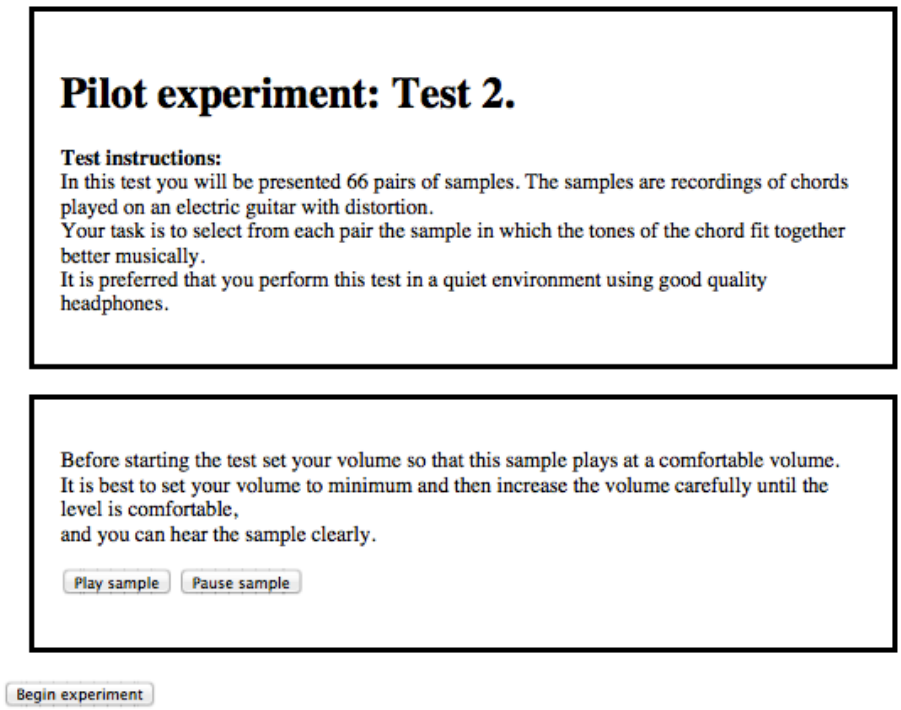
A1.3. Test 1 instructions



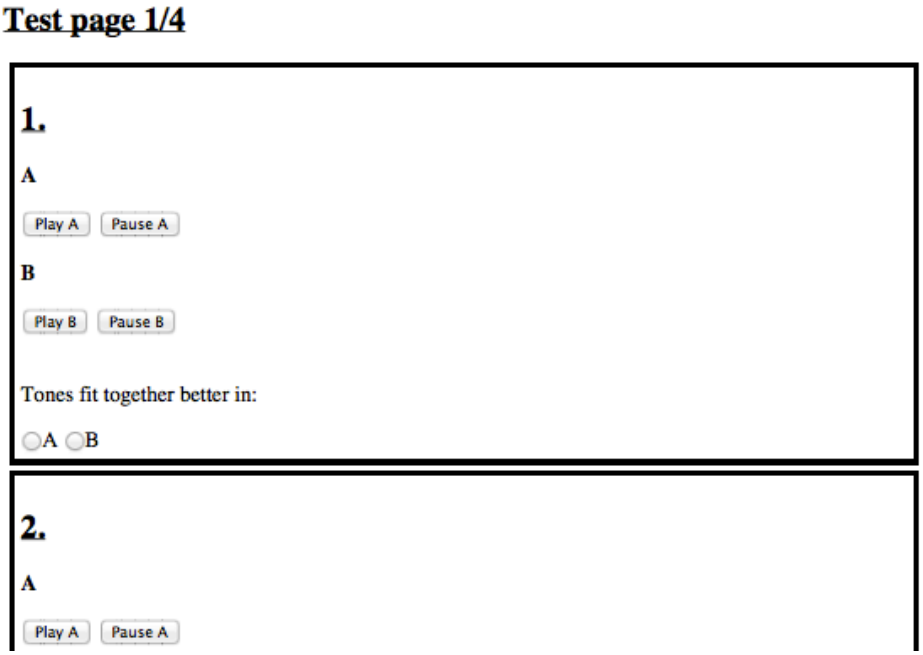

A1.4. Test 1 sample screenshot



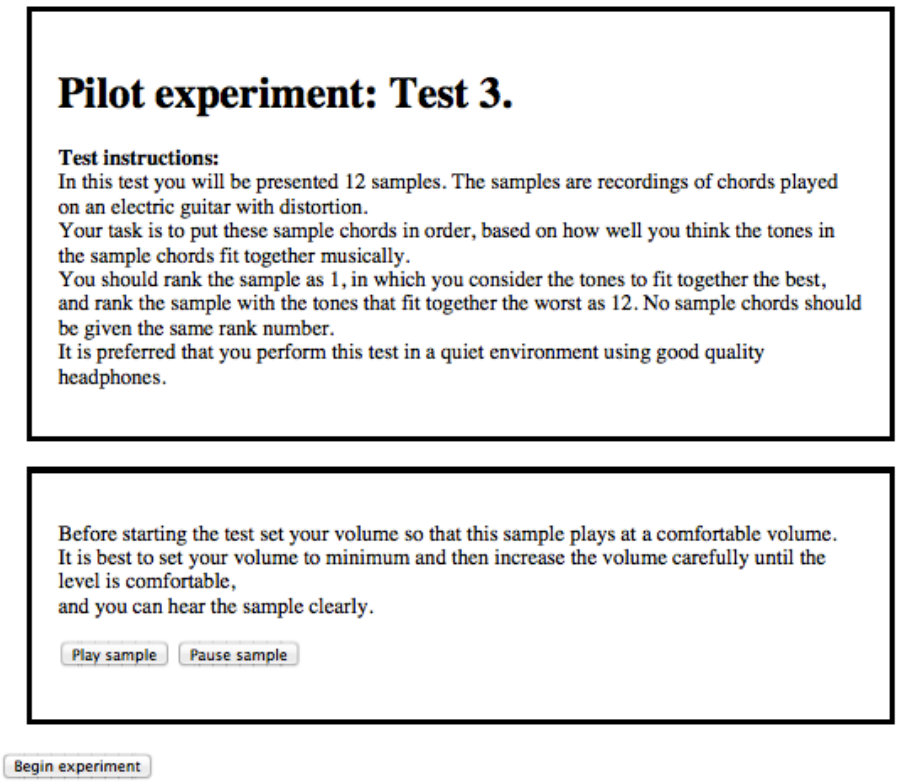
A1.5. Test 2 instructions



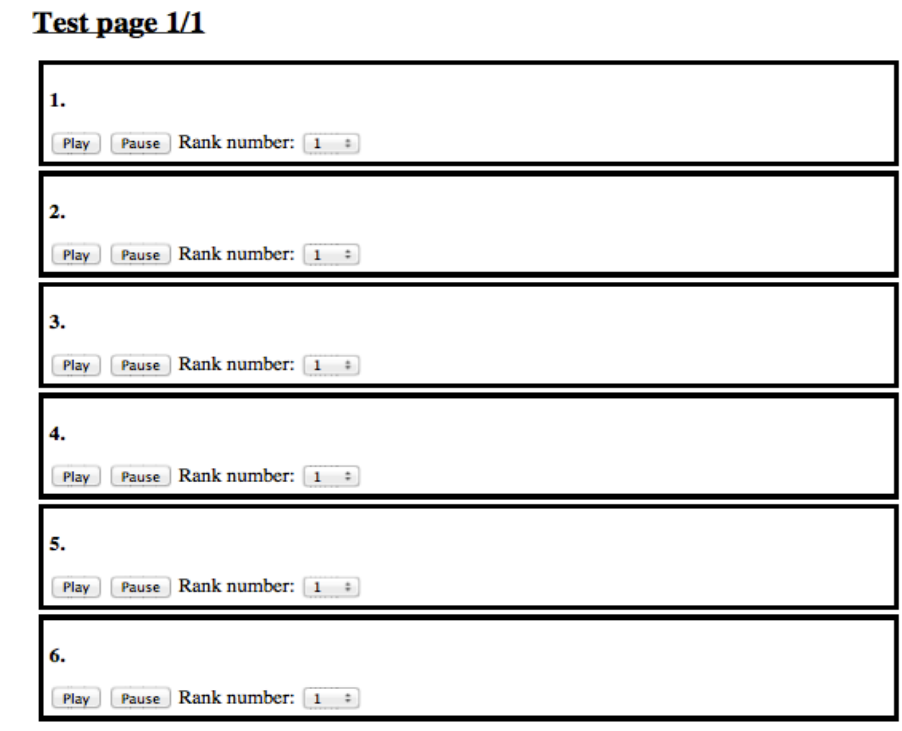

A1.6. Test 2 sample screenshot



A1.7. Test 3 instructions




A1.8. Test 3 sample screenshot



A1.9. Questionnaire at the end of each listening test

(*example from the end of test 1*)

**Questions on test 1.**

1. During the test did you feel like quitting the test?
- Yes
- No

2. Did you move on to these questions without completing the test (for example without listening to all of the samples)?
- Yes
- No

3. Did you ever start the test and quit it before completion, or did you complete it the first time you started it?
- I quit the test multiple times before completing it.
- I quit the test once before completing it.
- I completed the test the first time I started it.
- I didn't complete the test.

4. Was it it easy to understand the task you were asked to do in the test?
- 1 (Not at all easy)
- 2
- 3
- 4
- 5 (Very easy)

5. Were the instructions in the beginning of the test clear?
- Yes
- No

6. If the instructions were not clear, how should they be improved?

7. How boring did you find the test?
- 1 (Not at all boring)
- 2
- 3
- 4
- 5 (Very boring)

8. Did you find listening to the samples frustrating?
- 1 (Not at all frustrating)
- 2
- 3

4
5 (Very frustrating)

9. Did you take breaks during the test?
Yes
No

10. Did you visit other websites during the test?
Yes
No

11. Had you found this test on an Internet forum, would you have taken part?
Yes
No

12. Had you found this test on an Internet forum, would you have completed it?
Yes
No

13. You can write comments on this test in the box below (in English or Finnish):

By pressing submit your e-mail client will open a new message that contains your answers and the correct recipient. You only have to click "Send".
Once you have clicked Submit and sent the e-mail containing the results you can close this browser window.

Submit

A1.10. Questionnaire on the whole pilot experiment

## Questionnaire on the tests.

Please answer the following questions about the listening tests.
You can answer the questions in English or Finnish.

1. Did you find the questions on personal information and musical background intrusive?
- ○ Yes
- ○ No

2. If you found the questions intrusive why?

3. Do you think that there should have been more questions about musical background? if so, what kind of questions were lacking? (Leave empty if the questions were sufficient.)

4. In which listening test was it the easiest to make judgements on the samples?
- ○ Test 1, rating from 1–7.
- ○ Test 2, comparing pairs.
- ○ Test 3, putting samples in order.

5. Which of the tests did you like the most?
- ○ Test 1, rating from 1–7.
- ○ Test 2, comparing pairs.
- ○ Test 3, putting samples in order.

6. Did you recognize intervals and use your knowledge of music theory in making judgments?
- ○ Yes
- ○ No

7. If you came across these tests on the Internet, which one would you be most likely to quit (ie. NOT complete)?
- ○ Test 1, rating from 1–7.
- ○ Test 2, comparing pairs.
- ○ Test 3, putting samples in order.

8. You can write other comments on the tests in the textarea below.
For example things you liked or disliked in the tests.

Once you have clicked Submit and sent the e-mail containing the results you can close this browser window.

Submit

Appendix A2: The results of the pilot experiment

## A2.1. Results of musical background questionnaire

**Musical background questionnaire results**

**1. Age of subjects**

| | |
|---|---|
| Range | 21–30 |
| Average | 25 |

**2. Weekly hours of listening to music**

| | |
|---|---|
| Range | 2–40 |
| Average | 16.1 |

**3. Listener type rating**

(1 = casual listener, 5 = enthusiastic listener)

| | |
|---|---|
| Average | 4.13 |

**4. Had received training (instrument or vocal)**

| | |
|---|---|
| Yes | 100.0% |
| No | 0.0% |

**5. Had training in music theory / ear training**

| | |
|---|---|
| Yes | 100.0% |
| No | 0.0% |

**6. Duration of musical training in years**

| | |
|---|---|
| Average | 11.00 |

**7. Interval recognition skills**

(1 = can't recognise at all,
5 = can recognise with very good accuracy)

| | |
|---|---|
| Average | 3.25 |

**8. Had perfect pitch**

| | |
|---|---|
| Yes | 12.5% |
| No | 87.5% |

**9. Time taken to fill in questionnaire on musical background**

| | |
|---|---|
| Average | 2min42sec |

Comments on musical background questionnaire

Some of the comments were written in Finnish and some were written in English. Translations from Finnish to English are in *italics.* All translations by Otso Björklund.

1. Kuunneltavia musiikintyylejä kysyttäessä valmiit vaihtoehdot saattavat ohjata vastaamista. Itse ainakin laitoin "other" kohtaan eniten.
   *When asking about what styles of music you listen to the ready categories might guide the answers. I put most of my listening in the "other" category.*
2. "How many hours do you spend listening to music weekly?" Tässä kysymyksessä arviot saattavat heitellä aika paljon. Itse kuuntelen musiikkia keskittymisen eri ääripäissä. Välillä musiikki on taustalla, enkä kiinnitä siihen huomiota, välillä kuuntelu jakaa huomioni jonkin toisen aktiviteetin välillä, ja joskus kuuntelen musiikkia tarkasti tekemättä mitään muuta. Vastasin kysymkseen jonkinlaisena kompromissina näistä kaikista kuuntelun tyyleistä.
   *"How many hours do you spend listening to music weekly?" In this questions the estimates can vary a lot. I listen to music in the extremities of concentration. Sometimes music is in the background and I don't pay much attention to it, sometimes listening divides my attention between it and some other activity, and sometimes I listen to music attentively without doing anything else. My answer to the question was some kind of a compromise between these different types of listening.*
3. Ovatko "music theory / ear training" aina välttämättä yhdessä? Itse olen opetellut suhteellisen pitkälle teoriaa, mutta korvan kehittäminen on ollut huomattavasti epäsäännöllisempää.
   *Are "music theory / ear training" necessarily alway together? I have studied theory to quite an advanced level, but ear training has been considerably more irregular for me.*
4. Maybe a question about the format of the musical training might be interesting, as in formal, non-formal, conservatory/music school vs. private lessons etc.

A2.2. Results of test 1

**Test 1 results**

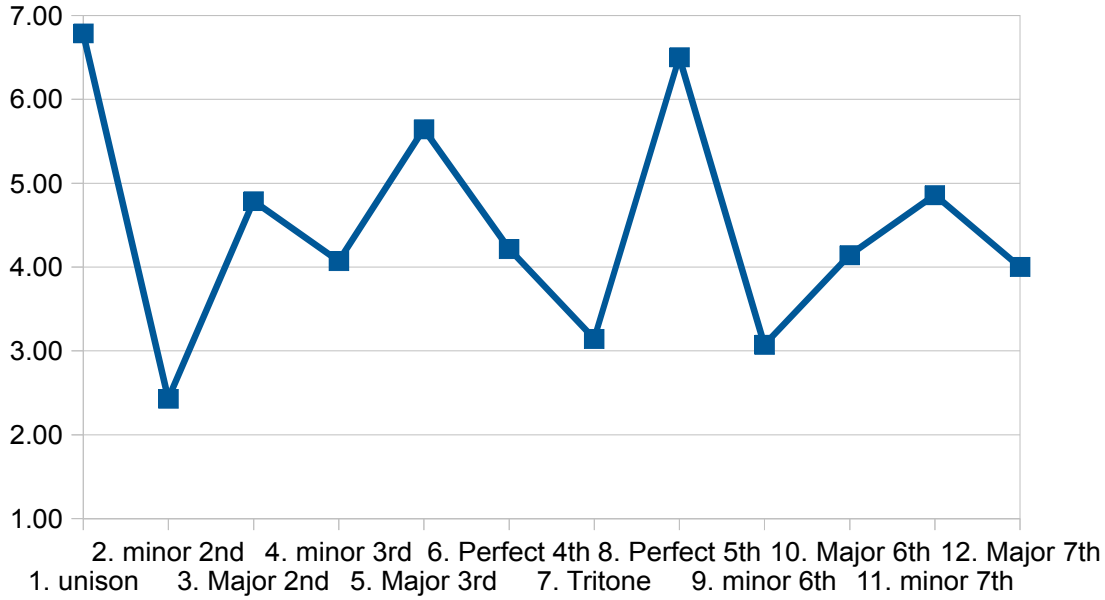
| Interval | Avg | STD |
|---|---|---|
| 1. unison | 6.79 | 0.43 |
| 2. minor 2nd | 2.43 | 1.28 |
| 3. Major 2nd | 4.79 | 1.37 |
| 4. minor 3rd | 4.07 | 1.21 |
| 5. Major 3rd | 5.64 | 1.01 |
| 6. Perfect 4th | 4.21 | 1.12 |
| 7. Tritone | 3.14 | 1.17 |
| 8. Perfect 5th | 6.50 | 0.52 |
| 9. minor 6th | 3.07 | 1.21 |
| 10. Major 6th | 4.14 | 1.35 |
| 11. minor 7th | 4.86 | 0.66 |
| 12. Major 7th | 4.00 | 1.36 |

Avg = Average
STD = Standard deviation

**Ordinal scale:**
*Most consonant*

| | |
|---|---|
| 1. | unison |
| 2. | Perfect 5th |
| 3. | Major 3rd |
| 4. | minor 7th |
| 5. | Major 2nd |
| 6. | Perfect 4th |
| 7. | Major 6th |
| 8. | minor 3rd |
| 9. | Major 7th |
| 10. | Tritone |
| 11. | minor 6th |
| 12. | minor 2nd |

*Least consonant*




Y-axis: average ratings, X-axis: interval.

A2.3. Results of test 2

**Test 2 results**

| **Interval** | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. | 9. | 10. | 11. | 12. | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. unison | – | 8 | 8 | 8 | 7 | 8 | 7 | 5 | 8 | 8 | 7 | 8 | 82 |
| 2. minor 2nd | 0 | – | 1 | 1 | 1 | 2 | 4 | 0 | 2 | 1 | 1 | 1 | 14 |
| 3. Major 2nd | 0 | 7 | – | 6 | 1 | 4 | 7 | 0 | 4 | 6 | 4 | 7 | 46 |
| 4. minor 3rd | 0 | 7 | 2 | – | 1 | 3 | 6 | 0 | 7 | 4 | 2 | 4 | 36 |
| 5. Major 3rd | 1 | 7 | 7 | 7 | – | 7 | 8 | 2 | 8 | 7 | 7 | 8 | 69 |
| 6. Perfect 4th | 0 | 6 | 4 | 5 | 1 | – | 7 | 0 | 6 | 4 | 2 | 5 | 40 |
| 7. Tritone | 1 | 4 | 1 | 2 | 0 | 1 | – | 0 | 6 | 1 | 2 | 1 | 19 |
| 8. Perfect 5th | 3 | 8 | 8 | 8 | 6 | 8 | 8 | – | 8 | 8 | 8 | 8 | 81 |
| 9. minor 6th | 0 | 6 | 4 | 1 | 0 | 2 | 2 | 0 | – | 2 | 2 | 3 | 22 |
| 10. Major 6th | 0 | 7 | 2 | 4 | 1 | 4 | 7 | 0 | 6 | – | 7 | 3 | 41 |
| 11. minor 7th | 1 | 7 | 4 | 6 | 1 | 6 | 6 | 0 | 6 | 1 | – | 5 | 43 |
| 12. Major 7th | 0 | 7 | 1 | 4 | 0 | 3 | 7 | 0 | 5 | 5 | 3 | – | 35 |
| | | | | | | | | | | | **Total:** | | **528** |

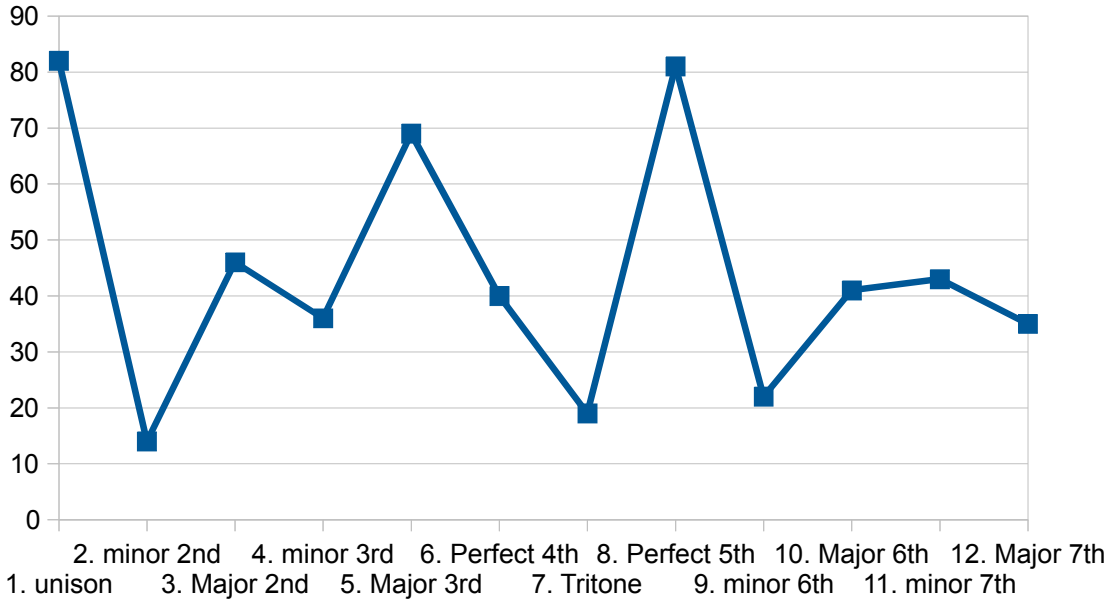
**Ordinal scale:**

*Most consonant*

| | |
|---|---|
| 1. | unison |
| 2. | Perfect 5th |
| 3. | Major 3rd |
| 4. | Major 2nd |
| 5. | minor 7th |
| 6. | Major 6th |
| 7. | Perfect 4th |
| 8. | minor 3rd |
| 9. | Major 7th |
| 10. | minor 6th |
| 11. | Tritone |
| 12. | minor 2nd |

*Least consonant*

Read: for each row, how many times it has been preferred to the columns
For example row 1, column 3 means how many times unison has been preferred to Major 2nd.




Y-axis: number of preferences, X-axis: interval

## A2.4. Results of test 3

**Test 3 results**

**Rankings for each subject (A to H)**

| **Interval** | A. | B. | C. | D. | E. | F. | G. | H. |
|---|---|---|---|---|---|---|---|---|
| 1. unison | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 |
| 2. minor 2nd | 10 | 12 | 12 | 12 | 3 | 11 | 9 | 11 |
| 3. Major 2nd | 8 | 4 | 7 | 5 | 7 | 10 | 6 | 4 |
| 4. minor 3rd | 4 | 8 | 9 | 7 | 12 | 4 | 4 | 7 |
| 5. Major 3rd | 3 | 5 | 3 | 6 | 4 | 3 | 3 | 3 |
| 6. Perfect 4th | 5 | 3 | 5 | 8 | 11 | 5 | 7 | 5 |
| 7. Tritone | 11 | 11 | 8 | 10 | 6 | 9 | 12 | 10 |
| 8. Perfect 5th | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| 9. minor 6th | 12 | 10 | 11 | 11 | 8 | 12 | 10 | 12 |
| 10. Major 6th | 6 | 6 | 6 | 4 | 9 | 8 | 8 | 9 |
| 11. minor 7th | 7 | 7 | 10 | 2 | 5 | 7 | 5 | 6 |
| 12. Major 7th | 9 | 9 | 4 | 9 | 10 | 6 | 11 | 8 |

Each column contains the rankings a subject has given to the samples

**Preference numbers (12 – rank)**

| **Interval** | A. | B. | C. | D. | E. | F. | G. | H. | Total |
|---|---|---|---|---|---|---|---|---|---|
| 1. unison | 11 | 11 | 10 | 9 | 11 | 11 | 11 | 11 | 85 |
| 2. minor 2nd | 2 | 0 | 0 | 0 | 9 | 1 | 3 | 1 | 16 |
| 3. Major 2nd | 4 | 8 | 5 | 7 | 5 | 2 | 6 | 8 | 45 |
| 4. minor 3rd | 8 | 4 | 3 | 5 | 0 | 8 | 8 | 5 | 41 |
| 5. Major 3rd | 9 | 7 | 9 | 6 | 8 | 9 | 9 | 9 | 66 |
| 6. Perfect 4th | 7 | 9 | 7 | 4 | 1 | 7 | 5 | 7 | 47 |
| 7. Tritone | 1 | 1 | 4 | 2 | 6 | 3 | 0 | 2 | 19 |
| 8. Perfect 5th | 10 | 10 | 11 | 11 | 10 | 10 | 10 | 10 | 82 |
| 9. minor 6th | 0 | 2 | 1 | 1 | 4 | 0 | 2 | 0 | 10 |
| 10. Major 6th | 6 | 6 | 6 | 8 | 3 | 4 | 4 | 3 | 40 |
| 11. minor 7th | 5 | 5 | 2 | 10 | 7 | 5 | 7 | 6 | 47 |
| 12. Major 7th | 3 | 3 | 8 | 3 | 2 | 6 | 1 | 4 | 30 |
| | | | | | | | **Total:** | | **528** |

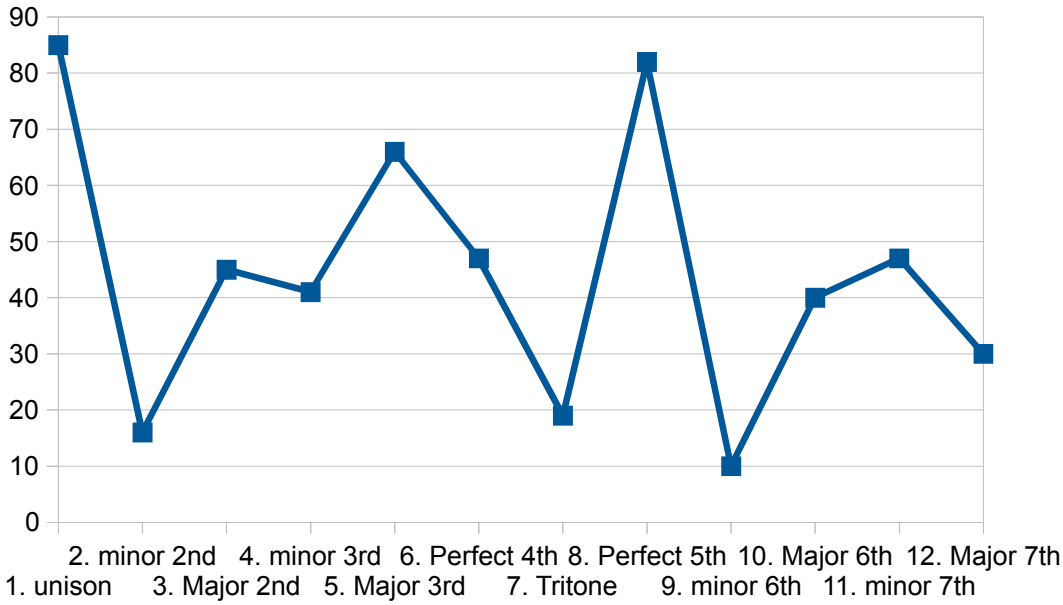
**Ordinal scale:**

*Most consonant*

| | | |
|---|---|---|
| 1. | unison | |
| 2. | Perfect 5th | |
| 3. | Major 3rd | |
| *4.* | *minor 7th* | *4. and* |
| *5.* | *Perfect 4th* | *5. are tied* |
| 6. | Major 2nd | |
| 7. | minor 3rd | |
| 8. | Major 6th | |
| 9. | Major 7th | |
| 10. | Tritone | |
| 11. | minor 2nd | |
| 12. | minor 6th | |

*Least consonant*

Rankings are mapped into the number of other samples a sample is preferred to. For example if a sample is ranked as 1 it is preferred to all of the other 11 samples, and if a sample is ranked 12 it is not preferred to any other sample (0 samples). Preference number is calculated using the following formula:
*Preference number = 12 – rank*




Y-axis: number of preferences, X-axis: interval

A2.5. Questionnaire results for test 1

**Test 1 questionnaire results**

**1. During the test did you feel like quitting the test?**

| | |
|---|---|
| Yes | 0.0% |
| No | 100.0% |

**2. Did you move on to these questions without completing the test (for example without listening to all of the samples)?**

| | |
|---|---|
| Yes | 0.0% |
| No | 100.0% |

**3. Did you ever start the test and quit it before completion, or did you complete it the first time you started it?**

| | |
|---|---|
| I quit the test multiple times before completing it. | 0.0% |
| I quit the test once before completing it. | 25.0% |
| I completed the test the first time I started it. | 75.0% |
| I didn't complete the test. | 0.0% |

**4. Was it easy to understand the task you were asked to do on the test?**

1 = Not at all easy, 5 = very easy

| | |
|---|---|
| Average | 3.63 |

**5. Were the instructions in the beginning of the test clear?**

| | |
|---|---|
| Yes | 62.5% |
| No | 37.5% |

**6. Comments on test instructions, see Comments on test 1.**

**7. How boring did you find the test?**

1 = Not at all boring, 5 = Very boring

| | |
|---|---|
| Average | 2.63 |

**8. Did you find listening to the samples frustrating?**

1 = Not at all frustrating, 5 = Very frustrating

| | |
|---|---|
| Average | 2.50 |

**9. Did you take breaks during the test?**

| | |
|---|---|
| Yes | 0.0% |
| No | 100.0% |

**10. Did you visit other websites during the test?**

| | |
|---|---|
| Yes | 12.5% |
| No | 87.5% |

**11. Had you found this test on an Internet forum, would you have taken part?**

| | |
|---|---|
| Yes | 62.5% |
| No | 37.5% |

**12. Had you found this test on an Internet forum, would you have completed it?**

| | |
|---|---|
| Yes | 37.5% |
| No | 62.5% |

| | |
|---|---|
| **Average time taken:** | 5min 25sec |

Comments on test 1

1. Vähän yksitoikkoinen testi. Joissakin kysymyksissä olisi myös hyvä olla vaihtoehto "ehkä".
   *Quite a monotonous test. In some questions it would be good to have an option "maybe".*
2. About the instructions:
   Luinkohan ensimmäiset ohjeet huonosti? Vasta kun pääsin kokeilemaan useita eri ääninäytteitä, rupesin miettimään, minkä perusteella minun olisi pitänyt arvioida kahden äänen sopivuutta. Olisiko minun pitänyt laittaa konsonoivat 7 ja dissonoivat 1? Huomasin, että jotkin näytteistä olivat huomattavasti dissonoivampia, mutta näkisin, että ne voisi arvioida sopivan hyvin yhteen riippuen musiikin kontekstista.
   *I wonder if I read the instructions poorly? Only after trying many different samples I started to think on what grounds should I have judged how well two tones fit. Should I have judged consonant ones 7 and dissonant ones 1? I noticed that some of the samples were considerably more dissonant, but I thought that they could be judged to fit well depending on a musical context.*
3. I've conflicting feelings about some of my answers, in the sense that I can hear several of the "bad" or non-fitting tone combinations being useful musically, as in "waiting for a resolution".
4. Yhdellä sivulla on aika monta samplea kerrallaan. Kaikkien järjestäminen johdonmukaisesti 1 - 7 asteikolla vaatii koko sample kasan kuuntelemista useampaan kertaan ja arvosanojen uudelleen jakelua joka voi olla melko turhauttavaa.
   *There were quite many samples on a single page. Putting all logically in order on the scale 1 – 7 required listening to quite a lot of samples many times and redistributing the ratings, which can be quite frustrating.*
5. About the instructions:
   I think I ranked all the samples rather high, because I'm not sure if it was about "did you like what you heard" or was it "are they more consonant or dissonant". I liked weird intervals and ranked them high, even if that was not the point.

6. Frustrating level 4, because:

   You should be able to hear the samples from beginning again. If you click pause, you will start at that point where you stopped which is most likely the end of the sample. Maybe put a button: replay, that would give the sample from the start.

7. About the instructions:

   Epäselvää oli, mitä piti kuunnella. On varsin subjetiivista sopiiko äänet hyvin yhteen vai ei, sehän on makuasia. Haettiinko tässä että onko dissonanssi vai konsonanssi? Jos sitä haettiin, pitäisi ilmaista selvästi asia.

   *It was unclear, what I was supposed to listen for. It's very subjective whether tones fit well together or not, it's a matter of taste. Was the idea to find out if something is a dissonance or a consonance? If that was the idea, it should be stated clearly.*

8. Miksi 7 eri vaihtoehtoa? Pitäisikö välivaihtoehdollekin olla joku nimi? Ymmärsin kohdat 1 ja 7 mutta kohdat 2, 3, 4, 5, 6 jäivät epäselviksi että mitä niissä haettiin.

   *Why 7 different options? Should the middle options also have some name? I understood 1 and 7, but with 2, 3, 4, 5, 6 it was unclear to me what their point was.*

## A2.6. Questionnaire results for test 2

**Test 2 questionnaire results**

**1. During the test did you feel like quitting the test?**

| | |
|---|---|
| Yes | 37.5% |
| No | 62.5% |

**2. Did you move on to these questions without completing the test (for example without listening to all of the samples)?**

| | |
|---|---|
| Yes | 0.0% |
| No | 100.0% |

**3. Did you ever start the test and quit it before completion, or did you complete it the first time you started it?**

| | |
|---|---|
| I quit the test multiple times before completing it. | 0.0% |
| I quit the test once before completing it. | 0.0% |
| I completed the test the first time I started it. | 100.0% |
| I didn't complete the test. | 0.0% |

**4. Was it easy to understand the task you were asked to do on the test?**

1 = Not at all easy, 5 = very easy

| | |
|---|---|
| Average | 4.13 |

**5. Were the instructions in the beginning of the test clear?**

| | |
|---|---|
| Yes | 75.0% |
| No | 25.0% |

**6. Comments on test instructions, see Comments on test 2.**

**7. How boring did you find the test?**

1 = Not at all boring, 5 = Very boring

| | |
|---|---|
| Average | 3.25 |

**8. Did you find listening to the samples frustrating?**

1 = Not at all frustrating, 5 = Very frustrating

| | |
|---|---|
| Average | 3.38 |

**9. Did you take breaks during the test?**

| | |
|---|---|
| Yes | 25.0% |
| No | 75.0% |

**10. Did you visit other websites during the test?**

| | |
|---|---|
| Yes | 25.0% |
| No | 75.0% |

**11. Had you found this test on an Internet forum, would you have taken part?**

| | |
|---|---|
| Yes | 50.0% |
| No | 50.0% |

**12. Had you found this test on an Internet forum, would you have completed it?**

| | |
|---|---|
| Yes | 12.5% |
| No | 87.5% |

| | |
|---|---|
| **Average time taken:** | 13min27sec |

Comments on test 2

1. 66 kohtaa tuntui hurjan paljolta. Lisäksi on aika yksitoikkoista kuunnella aina A-pohjaisia sointuja. Luulen, että tämä myös vaikuttaa kuulemiseen, sillä olen varma, että vastasin joissakin kohdissa ristiin. Tämän jälkeen kukaan ei kyllä enää usko että mulla on absoluuttinen korva (epäilen sitä tällä hetkellä myös itse). Ylipäänsä sävelien yhteissoinnin laatu lienee aika subjektiivista.
   *66 trials felt like a lot. It's quite monotonous to always listen to A-based chords. I think that this also affects hearing, because I'm sure in some trials I answered inconsistently. After this no-one will think that I have perfect pitch (I also doubt it myself at the moment). Overall the quality of simultaneously playing pitches is probably quite subjective.*
2. About the instructions:
   Sama kuin testissä 1. Äänien sopivuuden arviointi oli välillä jokseenkin haastavaa. Joissain tapauksissa toinen vaihtoehto olisi todennäköisesti ollut sopivampi.

*Same as in test 1. Judging the fit of tones was sometimes fairly challenging. In some cases the other option could have probably been better.*

3. Sama kuin testissä 1: kysymyksessä 11 voisi olla "maybe". Vastaukseni varmaan riippuisi foorumista ja siitä kuinka testiä mainostetaan.
   *Same as in test 1: question 11 could also have the option maybe. My answers would possibly depend on the Internet forum and how it was being advertised.*
4. The test is a bit on the long side, I find my concentration slipping after a while, leading to repeated listens of the samples.
5. One bug found:
   If the the same sample is in two tests, and you pause the first one and move on, when you press the "Play A/B" -button the next time, the sample will not start from the beginning, but it will start from the position paused instead.
6. About the instructions:
   Mitä tarkoitetaan sillä että "fit musically together"?
   *What is meant by "fit musically together"?*

## A2.7. Questionnaire results for test 3

**Test 3 questionnaire results**

**1. During the test did you feel like quitting the test?**

| | |
|---|---|
| Yes | 0.0% |
| No | 100.0% |

**2. Did you move on to these questions without completing the test (for example without listening to all of the samples)?**

| | |
|---|---|
| Yes | 0.0% |
| No | 100.0% |

**3. Did you ever start the test and quit it before completion, or did you complete it the first time you started it?**

| | |
|---|---|
| I quit the test multiple times before completing it. | 0.0% |
| I quit the test once before completing it. | 0.0% |
| I completed the test the first time I started it. | 100.0% |
| I didn't complete the test. | 0.0% |

**4. Was it easy to understand the task you were asked to do on the test?**

1 = Not at all easy, 5 = very easy

| | |
|---|---|
| Average | 4.38 |

**5. Were the instructions in the beginning of the test clear?**

| | |
|---|---|
| Yes | 100.0% |
| No | 0.0% |

**6. Comments on test instructions, see Comments on test 3.**

**7. How boring did you find the test?**

1 = Not at all boring, 5 = Very boring

| | |
|---|---|
| Average | 2.38 |

**8. Did you find listening to the samples frustrating?**

1 = Not at all frustrating, 5 = Very frustrating

| | |
|---|---|
| Average | 2.75 |

**9. Did you take breaks during the test?**

| | |
|---|---|
| Yes | 12.5% |
| No | 87.5% |

**10. Did you visit other websites during the test?**

| | |
|---|---|
| Yes | 25.0% |
| No | 75.0% |

**11. Had you found this test on an Internet forum, would you have taken part?**

| | |
|---|---|
| Yes | 75.0% |
| No | 25.0% |

**12. Had you found this test on an Internet forum, would you have completed it?**

| | |
|---|---|
| Yes | 62.5% |
| No | 37.5% |

| | |
|---|---|
| **Average time taken:** | 7min21sec |

Comments on test 3

1. Näytteiden määrä oli sopiva, hermot eivät tällä kertaa menneet. Näytteiden järjestäminen tuntui myös kiinnostavalta kuin yksittäisten äänten/parin näytteen vertailu.
   *The number of samples was suitable, I didn't lose my nerve this time. Ordering samples also felt more interesting than judging single tones/pairs of samples.*
2. Sivulla, jossa näytteet asetetaan järjestykseen, voisi olla vielä ohjeet, että 1 tarkoittaa sopivinta ja 12 epäsopivinta.
   *On the page where you put the samples in order could also be instructions that 1 means most fitting and 12 least fitting.*
3. The user interface is very complicated. I had to search for a while for the number that was missing (and had two of some kind). Also it would be nice to order the elements based on your rating so it would be easier to compare in the end. Now you had to jump around, since at least I wanted to justify my order so I listened the my ordering and jumbled them around quite a bit, and now you have to scroll up and down to find where the next number is. Maybe some kind of drag'n'drop boxes UI that would visualize the order easily would be better. While I liked the idea of this test the best, the implementation made it the worst/hardest to complete.
4. Käyttöliittymä oli vaikea kun piti monta kertaa kuunnella kaikki äänitteet ja skrollata ylös ja alas sivua, että muistaa mikä numero milläkin on. Olisi helpompi, jos olisi liikuteltavia paloja, jotka voisi laittaa järjestykseen eikä tällainen numerointi.
   *The user interface was difficult, because you had to listen to all of the samples and scroll up and down the page to remember which number you had assigned to which sample. It would be easier to have movable boxes which you could put in order instead of having this kind of numbering.*

A2.8. General questionnaire results

**General questionnaire results**

**1. Did you find the questions on personal information and musical background intrusive?**

| | |
|---|---|
| Yes | 0.0% |
| No | 100.0% |

**4. In which listening test was it the easiest to make judgements on the samples?**

| | |
|---|---|
| Test 1 | 0.0% |
| Test 2 | 50.0% |
| Test 3 | 50.0% |

**5. Which of the tests did you like the most?**

| | |
|---|---|
| Test 1 | 0.0% |
| Test 2 | 25.0% |
| Test 3 | 75.0% |

**6. Did you recognize the intervals and use your knowledge of music theory in making judgements?**

| | |
|---|---|
| Yes | 87.5% |
| No | 12.5% |

**7. If you came across these tests on the Internet, which one would you be most likely to quit?**

| | |
|---|---|
| Test 1 | 0.0% |
| Test 2 | 62.5% |
| Test 3 | 37.5% |

General comments on the pilot experiment

1. Kuuntelunäytteiden määrä varsinkin kohdassa 2 oli mielestäni turhan suuri.
   *The number of audio samples especially in test 2 was too large in my opinion.*
2. Kontekstin puuttumisen takia olisin saattanut jättää vastaamatta. Voisiko "fit together better musically" tarkentaa hieman ilman että se vaikuttaisi vastauksiin?
   *Due to the lack of context I might have not answered. Could "fit together better musically" be clarified a bit without it affecting the ratings?*

3. Pidin eniten testistä 3, jossa ne laitettiin järjestykseen, mutta asettelussa voisi pyrkiä siihen, ettei sivulla tarvitsisi scrollata. Omalla ruudullani homman olisi voinut helposti jakaa kahteen palstaan, ja silti olisi jäänyt tyhjää tilaa.
   *I preferred test 3, in which they were put in order, but you could try to make such a layout that scrolling wouldn't be necessary. On my screen the thing could have been divided into two columns, and there still would have been excess space on the screen.*
4. Regarding question no 6. I wouldn't say I used knowledge of theory in the sense that "a tritone is dissonant" but more in the sense, that I know (or have been taught ;) ) that certain intervals "distort" better, so when I hear a fifth, I know it's going to "work", and that might cloud my judgement...
5. Comparing pairs oli ihan hyvä testi, mutta aivan liian pitkä. Olisin jaksanut tehdä siitä hyvin 1 tai 2 sivua, mutta ei missään tapauksessa 4. Se oli kuitenkin vähemmän epämääräinen, kun testi 1, jossa oli vaikea ymmärtää, miksi rating oli 1–7.
   *Comparing pairs was a pretty good test, but it was far too long. I could have completed one or two pages of it, but not 4 in any case. It was still less ambiguous than test 1, in which I had difficulty understanding why the rating was 1–7.*